2019

Longitudinal And Cross-Country Measurement Invariance Of The Pisa Home Possessions Scale

Selene S. Lee
University of Pennsylvania, selenelee0525@gmail.com

Longitudinal And Cross-Country Measurement Invariance Of The Pisa Home Possessions Scale

Abstract
Measuring socioeconomic status (SES) is very important in educational research, as researchers often use this information to contextualize the results of an assessment or to control for SES when analyzing the relationship between academic achievement and other variables. However, any cross-country comparisons using SES data from international large-scale assessments, such as the Programme for International Student Assessment (PISA), should be preceded by a careful examination of the psychometric properties of the scale used to measure SES, an issue which is rarely addressed by researchers. The current study aims to fill the gaps in this field of research by analyzing the longitudinal and cross-country measurement invariance of the PISA home possessions scale, a 25-item scale which measures household wealth, one of the three components used to measure SES in PISA. Using multiple group concurrent calibration with partial invariance constraints, the study found that four items in the scale, all related to technology, functioned differently across the PISA cycles. It also found that some items (i.e., bathroom, classic literature, poetry books, and TV) functioned differently across the participating countries when used to measure family wealth. The overall level of misfit found in the scale was not associated with the country's GDP per capita, while some evidence suggested that it may be associated with the region in which the country was located and sociocultural factors (which were partially captured by the language in which students took the assessment). Compared to the original home possessions scores obtained from the public dataset, the new home possessions scores generated with the method used in the study were found to be a more comparable measure of SES across countries, while the accuracy of the scores as a measure of SES within countries was improved in most cycles. The study also found validity evidence supporting the use of the new home possessions scores as a measure of SES. The results of this study can help improve the PISA home possessions scale, so it can continue to provide valuable information to researchers and policy makers on SES over the PISA cycles and across the countries that participate in PISA.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Education

First Advisor
Robert F. Boruch

Keywords
Home possessions scale, International large-scale assessment, Measurement invariance, PISA, Psychometrics, SES

LONGITUDINAL AND CROSS-COUNTRY MEASUREMENT INVARIANCE OF THE PISA HOME POSSESSIONS SCALE

Selene Sunmin Lee

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

_____

Robert Boruch, University Trustee Chair Professor
of Education and Statistics

Graduate Group Chairperson

_____

J. Matthew Hartley, Professor of Education

Dissertation Committee

Paul A. McDermott, Professor of Education

Michael J. Rovine, Senior Fellow

Matthias von Davier, Distinguished Research Scientist, NBME

LONGITUDINAL AND CROSS-COUNTRY MEASUREMENT INVARIANCE OF THE PISA HOME POSSESSIONS SCALE

COPYRIGHT

2019

Selene Sunmin Lee

DEDICATION

To my husband, Sangbaek, for all the personal sacrifices he made while I pursued my PhD.

ACKNOWLEDGMENTS

I would not be here today without the help of so many people during the past four years. First, I would like to thank my advisor, Dr. Robert Boruch, for guiding me and supporting me throughout my graduate studies, and for reminding me not to lose my sense of humor even during difficult times. His feedback on earlier drafts of my dissertation has made it much stronger. I would also like to thank Dr. Paul McDermott whose teachings on educational measurement inspired me to change the course of my studies. His classes, at times challenging, have made me into a well-rounded researcher, while his constant encouragement and trust in me has made me into a more confident one. Many thanks to Dr. Michael Rovine who first introduced me to the world of measurement invariance. I am grateful for the numerous hours he spent with me to answer questions about my dissertation and the papers we read for the QM reading group. Also, I would like to thank Christine Lee for helping me with all the administrative work, both big and small, and for providing me with moral support for the past four years.

In addition, I would like to express my sincere gratitude for Dr. Matthias von Davier who agreed to be on my dissertation committee, even though he barely knew me when I asked. His guidance was invaluable when I was formulating the research questions and whenever I had questions about the methodology. I am also grateful for my internship mentors at ETS last summer, Dr. Lale Khorramdel and Dr. Kentaro Yamamoto, for teaching me how to assess measurement invariance using PISA data and for making me believe that it was not crazy to change my dissertation topic a few months before my dissertation proposal.

Last but not least, I would like to thank my husband, Sangbaek, for supporting me from afar during the first three years of my graduate studies and for relocating to Philadelphia last year to take over the housework while I worked on my dissertation. You are the most patient person I have ever met, and I will forever be grateful for all the personal sacrifices you made for me.

ABSTRACT

LONGITUDINAL AND CROSS-COUNTRY MEASUREMENT INVARIANCE OF THE PISA HOME POSSESSIONS SCALE

Selene Sunmin Lee

Robert Boruch

Measuring socioeconomic status (SES) is very important in educational research, as researchers often use this information to contextualize the results of an assessment or to control for SES when analyzing the relationship between academic achievement and other variables. However, any cross-country comparisons using SES data from international large-scale assessments, such as the Programme for International Student Assessment (PISA), should be preceded by a careful examination of the psychometric properties of the scale used to measure SES, an issue which is rarely addressed by researchers. The current study aims to fill the gaps in this field of research by analyzing the longitudinal and cross-country measurement invariance of the PISA home possessions scale, a 25-item scale which measures household wealth, one of the three components used to measure SES in PISA. Using multiple group concurrent calibration with partial invariance constraints, the study found that four items in the scale, all related to technology, functioned differently across the PISA cycles. It also found that some items (i.e., bathroom, classic literature, poetry books, and TV) functioned differently across the participating countries when used to measure family wealth. The overall level of misfit found in the scale was not associated with the country's GDP per capita, while some evidence suggested that it may be associated with the region in which the country

was located and sociocultural factors (which were partially captured by the language in which students took the assessment). Compared to the original home possessions scores obtained from the public dataset, the new home possessions scores generated with the method used in the study were found to be a more comparable measure of SES across countries, while the accuracy of the scores as a measure of SES within countries was improved in most cycles. The study also found validity evidence supporting the use of the new home possessions scores as a measure of SES. The results of this study can help improve the PISA home possessions scale, so it can continue to provide valuable information to researchers and policy makers on SES over the PISA cycles and across the countries that participate in PISA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ACROYNYMS

2PL model	Two-parameter logistic model
DHS	Demographic Health Surveys
DIF	Differential item functioning
ESCS	Economic, social and cultural status
GDP	Gross Domestic Product
GNI	Gross National Income
GPCM	Generalized partial credit model
HDI	Human Development Index
ICC	Item characteristic curve
IEA	Int'l Association for the Evaluation of Educational Achievement
ILO	International Labour Organization
IRT	Item response theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index of Occupational Status
MD	Mean deviation
MIRT	Multidimensional item response theory
OECD	Organization for Economic Co-operation and Development
PCA	Principal component analysis
PCM	Partial credit model
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RMSD	Root mean square deviation
SES	Socioeconomic status

TERCE	Third Regional Comparative and Explanatory Study
TIMSS	Trends in International Mathematics and Science Studies
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	U.S. Agency for International Development

CHAPTER 1 – INTRODUCTION

Importance of Measuring Socioeconomic Status

In international comparative research in education, it is important to collect reliable and valid information on students' socioeconomic status (SES), as researchers often use this information to contextualize the results of an assessment or to control for SES when analyzing the relationship between academic achievement and other variables. While most of the research on SES and students' academic achievement have been conducted in developed countries, international large-scale assessments such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS) have made it possible to conduct such research in a wide range of countries as well as to make cross-country comparisons. For example, in recent years, the Organization for Economic Co-operation and Development (OECD) has published several reports based on cognitive scores and SES information collected from students in countries that participated in PISA. One report, focusing on students with low SES, compared students who received high scores on PISA with those who received low scores (OECD, 2011). The study found that the former had more self-confidence in their academic abilities and also spent more time in class, leading to recommendations that schools should foster the self-confidence of students from low SES and ensure that they spend sufficient time in class. Another report found a strong correlation between student performance in PISA and early career outcomes and suggested that policies that focus on

equity in educational achievement may be able to foster social mobility in the long term (OECD, 2018). However, as noted by Rutkowski and Rutkowski (2013), any cross-country comparisons using SES data from international large-scale assessments should be preceded by a careful examination of the psychometric properties of the scale used to measure SES, a topic which is rarely addressed by researchers.

The current study is designed to fill the gaps in this area of research. It is especially timely, given the increasing interest not only among researchers but also among education policy makers in analyzing the relationship between SES and educational outcomes, following the increased focus on equity in the global education agenda. This focus on educational equity is clearly highlighted in the Sustainable Development Goals, a set of 17 goals launched by the United Nations in 2015, of which the fourth goal is to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations, 2015). Educational equity is also highlighted in the Education 2030 Framework for Action, the framework outlining the strategies for achieving the fourth goal of the Sustainable Development Goals, which recommends policies to address the uneven distribution of learning opportunities and outcomes across regions, households, ethnic, and socioeconomic groups (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2016).

Although there is some disagreement on the conceptual meaning of SES, a general consensus emerged among researchers by the 1980s that it should be measured by a composite of several indicators (Brese & Mirazchiyski, 2013). The most widely used indicators to measure SES are income, education, and occupation, as proposed by

Duncan, Featherman, and Duncan in 1972 (Brese & Mirazchiyski, 2013; Rutkowski & Rutkowski, 2013; Sirin, 2005). Among the three indicators, this research will focus on family income, as it is the only indicator of SES measured with a scale in PISA, making it possible for its psychometric properties to be evaluated.

The remainder of this chapter will present a literature review on measuring family income with home possessions, followed by the research questions for the five studies included in the research. Chapter 2 will cover the methodology of each study, and the results will be presented in Chapter 3. Lastly, Chapter 4 will address the significance of the studies as well as its limitations.

Difficulties of Measuring Family Income Directly and Indirectly

Many researchers have written about the difficulties of measuring family income directly (Brese & Mirazchiyski, 2013; Filmer & Pritchett, 2001; Tourangeau & Yan, 2007; Vyas & Kumaranayake, 2006). This is because survey questions about income are often regarded as intrusive, leading to a high non-response rate regardless of the respondent's level of income (Brese & Mirazchiyski, 2013; Tourangeau & Yan, 2007). As a case in point, PISA 2015 included an item in the parent questionnaire regarding the family's annual income. This item had six response categories (with different response categories for each country), and it was also accompanied by a note ensuring respondents that their response would be kept strictly confidential (OECD, 2014a). However, among the 72 countries that participated in PISA 2015, only 18 countries opted to administer the parent questionnaire (OECD, 2017, p. 332), and among these countries, two countries

decided to exclude this item. Among the countries that administered this item, the response rate was only 59%. As a result, information on family income was available for only 13% of the students that participated in PISA 2015.

In the absence of reliable information on family income, home possessions (i.e., assets owned by the household) have often been used as a proxy (Filmer & Scott, 2008; Montgomery, Gragnolati, Burke, & Paredes, 2000; Vyas & Kumaranayake, 2006). According to Filmer and Pritchett (2001), one of the first researchers to use home possessions as a proxy for family income, home possessions are at least as reliable as conventional measures of family income in predicting educational outcomes. This is because the educational decisions of households are usually based on their long-term economic situation, which is well-reflected in their home possessions status. Another advantage of collecting information on home possessions is that it does not depend on the exchange rate, making it easier to make comparisons across countries (Brese & Mirazchiyski, 2013). For these reasons, home possessions have been used as a proxy for family income in many studies and surveys, especially in contexts in which it is difficult to collect reliable information on family income, such as when the study is conducted in developing countries (Vyas & Kumaranayake, 2006) or when the subjects are children (Brese & Mirazchiyski, 2013).

However, there are challenges to using home possessions as a proxy for family income. First, home possessions are more likely to be a measure of family wealth, which refers to the stock of family resources at a certain point in time, so it may not be an accurate measure of family income, which refers to the flow of family resources over an

interval of time. For this reason, in this research, home possessions will be considered to be a measure of family wealth instead of family income. Another problem is that if respondents are asked about the exact number of an item that is owned by their household, their response may be positively correlated with the number of people in their household, producing a spuriously high estimate of household wealth for larger families. Considering that the average family size of a country is influenced by sociocultural factors, estimates of household wealth that are generated with this method may not be comparable across countries. Lastly, ownership of an item does not convey information about the quality of the item that is owned (Falkingham & Namazie, 2002), how accessible the item is in a country due to economic and logistical reasons, or how valued it is due to sociocultural reasons (Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004). This touches upon the issue of measurement invariance across countries, which will be explored in detail later in this study.

How Home Possessions are Surveyed in International Assessments and International Household Surveys

Without a consensus among researchers on how family wealth should be measured with information on home possessions (Vyas & Kumaranayake, 2006), it is not surprising that there is no widely used scale to collect information on respondents' home possessions. The following section explains how home possessions are surveyed in several major international large-scale assessments and international household surveys.

PISA is an international large-scale assessment coordinated by the OECD to assess the knowledge and skills of 15-year-old students in reading, math, and science, and it has been administered every three years since 2000 (“About PISA,” n.d.). In PISA 2015, the most recent cycle of PISA for which data are publicly available, the student questionnaire included 22 items regarding students’ home possessions. This scale will be presented in detail later.

TIMSS is another major international large-scale assessment which is coordinated by the International Association for the Evaluation of Educational Achievement (IEA) to measure 4th- and 8th-grade students’ achievement in math and science every four years (“TIMSS overview,” n.d.). In the most recent cycle of TIMSS in 2015, the student questionnaire included eight items regarding students’ home possessions (IEA, 2014). Four of these items overlapped with the home possessions scale of PISA (i.e., desk to study at, room of your own, link to the internet, and the number of books), three items were similar (i.e., computer or tablet, own mobile phone, and digital information devices), and only one item was unique to TIMSS (i.e., gaming system). Also, TIMSS allowed countries to include up to four country-specific items, one more item than in PISA. However, unlike PISA, TIMSS did not use these items to produce a single score for each household representing their household wealth.

PIRLS is another major international large-scale assessment coordinated by the IEA to measure 4th-grade students’ reading skills every five years (“PIRLS overview,” n.d.). In the latest cycle of PIRLS in 2016, the student questionnaire included five items regarding students’ home possessions (IEA, 2015). Four of these items overlapped with

the home possessions scale of PISA (i.e., desk to study at, room of your own, link to the internet, and the number of books), and one item was similar (i.e., computer or tablet). As in TIMSS, countries were allowed to include up to four country-specific items, but a single score representing the household wealth of each household was not produced.

The Demographic Health Surveys (DHS) is an international household survey supported by the U.S. Agency for International Development (USAID) to provide nationally representative information on the health, nutrition, and population of low- and middle-income countries ("DHS overview," n.d.). DHS collects information on the housing characteristics and home possessions of the households that are surveyed, including the main material of the floor, roof, and wall; characteristics of the toilet facility, handwashing facility, cooking facility, and heating facility; source of water and light; connection to electricity, cable services, internet, and fixed telephone line; ownership of livestock, vehicles, a bank account, and other household items. For each country, a principal component analysis (PCA) is conducted with this information, with separate component loadings estimated for urban and rural areas (Rutstein, 2008). The household's score on the first principal component is considered to represent the wealth of the household. However, unlike PISA, these scores are only used to rank the wealth of households within a country (Rutstein & Johnson, 2004), not to make comparisons across countries. Thus, the items and response categories in the DHS surveys are not identical across countries, nor are the surveys conducted simultaneously in the participating countries.

Challenges of Ensuring Measurement Invariance across Countries

When the aim of surveying home possessions is to produce scores of family wealth that are comparable across countries, it is necessary to first establish the extent to which measurement invariance of the scale can be assumed across countries. Measurement invariance, in this context, implies that the relationship between the ownership of an item and household wealth (i.e., the latent variable) does not depend on the country in which the scale is administered. As explained above, this may not hold if the accessibility of an item varies across countries due to economic and logistical reasons, or the value of an item varies due to sociocultural reasons (Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004). When measurement invariance is not established, household wealth scores generated from the scale cannot be meaningfully compared across countries (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014).

Traditionally, measurement invariance has been tested using multiple group confirmatory factor analysis (Jöreskog, 1971; Meredith, 1993). In this approach, the observed score for an item is modeled as a linear function of the factor score, and increasingly strict constraints are placed on the parameters to establish higher levels of measurement invariance. Configural invariance, the lowest level of invariance, requires the factor structure to be identical across groups; metric (or weak) invariance requires the factor loadings to also be identical across groups; and scalar (or strong) invariance requires the intercept of the regression equations to also be identical across groups. However, when many groups and items are included in the analysis, for example, when the analysis is conducted with data from international large-scale assessments, scalar

invariance is rarely achievable in practice (Davidov et al., 2014; Davidov, Muthén, & Schmidt, 2018; Marsh et al., 2018). For example, in a study by Sandoval-Hernandez, Rutkowski, Matta, and Miranda (2019), only configural invariance could be established for the Economic, Social and Cultural Status scale of PISA 2015 which was administered to 72 countries, as well as for the Home Educational Resources scale of TIMSS 2015 which was administered to 44 countries. For the Family Socioeconomic and Cultural Status scale of the Third Regional Comparative and Explanatory Study (TERCE), a learning assessment administered in 16 Latin American countries in 2013, only metric invariance could be established.

Due to the difficulties of establishing scalar invariance in practice, in recent years, several alternative methods have been proposed to assess the measurement invariance of scales when many groups and items are included in the analysis. These methods assume that meaningful comparisons can still be made across groups when there are some violations that threaten the equality of the measurement model across groups. Therefore, weaker constraints are imposed on the model compared to multiple group confirmatory factor analysis (Davidov et al., 2014). One such method is based on Bayesian structural equation modeling (Muthén & Asparouhov, 2012) which allows for small differences in the factor loadings and intercepts across groups. In this model, these differences are considered to be variables for which the distribution can be described by a substantive prior distribution centered around zero (Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015). In a study by Cieciuch, Davidov, Algesheimer, and Schmidt (2018), this method was used to establish approximate measurement invariance

for six cycles of the European Social Survey administered in 15 countries between 2002 and 2013.

Another method that has been proposed is the alignment method (Muthén & Asparouhov, 2014) which is implemented in a multiple group confirmatory factor analysis or structural equation modeling framework. In this method, only the parameters that have large differences across countries are relaxed, resulting in a solution that has a few non-invariant parameters with large differences instead of many non-invariant parameters with small differences (Davidov et al., 2018). Munck, Barber, and Torney-Purta (2018) used this method to establish approximate measurement invariance for the 1999 Civic Education Study and the 2009 International Civics and Citizenship Education Study administered in 18 European countries.

The alignment method described above is equivalent to multiple group concurrent calibration with partial invariance constraints which establishes approximate invariance across populations in the context of multidimensional item response theory (MIRT). The basic form of this procedure allows some item parameters to vary if large deviations of item functions are detected (Glas & Verhelst, 1995; von Davier & von Davier, 2007; Yamamoto, 1998; Yamamoto & Mazzeo, 1992). Specifically, for each item and group, item fit statistics are computed to quantify the discrepancy between the observed item characteristic curve (ICC) for the group and the model-based ICC estimated with data from all groups. When the item fit statistic is over a certain threshold, differential item functioning (DIF) is assumed for the group, which means that measurement invariance cannot be established for the item. A fully automated algorithm for this method was

developed for the analysis of the Programme for the International Assessment of Adult Competencies (PIAAC) data (Glas & Jehangir, 2014; Oliveri & von Davier, 2011; Xu & von Davier, 2008, Yamamoto, Khorramdel, & von Davier, 2013). This method was also used to assess the measurement invariance of the items in the cognitive assessment and background questionnaire for PISA 2015 (OECD, 2017, p. 295) and PISA 2018. Details on this method will be presented later, as it is the method that was used in the current research.

Research Questions

The present study assesses the measurement invariance of PISA's home possessions scale which is one of the three components, along with parents' education and parents' occupation, used to measure students' SES in PISA. This research is very timely, given the increasing interest among researchers and policy makers in analyzing educational equity, both longitudinally and across countries, as well as the increasing heterogeneity of countries that are participating in PISA. The specific research questions for the current study are described below.

Research question 1: Which items in the PISA home possessions scale demonstrate measurement invariance over multiple PISA cycles? This research question was addressed in Study 1 which examined the measurement invariance of the items in the PISA home possessions scale over the six PISA cycles conducted in 2000, 2003, 2006, 2009, 2012, and 2015. For each item, the default was to constrain the item parameters to be equal across the cycles. Subsequently, cycles for which the observed

ICC exhibited substantial misfit with the total sample ICC were allowed to estimate their own item parameters, essentially creating a partial scalar invariance model (Byrne, Shavelson, & Muthén, 1989). This study made it possible to analyze whether and how the item parameters shifted across the cycles, since the item parameters for all cycles were estimated on a common scale.

Research question 2: Which items in the PISA home possessions scale demonstrate measurement invariance across the country-by-language groups? This research question was addressed in Study 2 which examined the measurement invariance of the items in the PISA home possessions scale across the 96 country-by-language groups that participated in PISA.[1] Using the final model of Study 1 as the initial model, the default was to constrain the item parameters to be equal across the country-by-language groups. Subsequently, country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC were allowed to estimate their own item parameters. This study made it possible to analyze whether and how the item parameters varied across the country-by-language groups, since the item parameters for all country-by-language groups were estimated on a common scale.

Research question 3: Are the new home possessions scores a more comparable measure of SES across countries than the original home possessions scores? To test whether the new home possessions scores (generated from the final model in Study 2) were a more comparable measure of SES across countries than the

[1] In Study 2, students within countries were grouped by the language in which they took the cognitive assessment. Languages that were used as the language of examination by at least 5% of the test takers in the country (using final student weights) were considered to be independent country-by-language groups.

original home possessions scores (obtained from the public dataset), in Study 3, PCAs were conducted using home possessions, parents' education, and parents' occupation – the three components used to measure students' SES in PISA. The analyses were conducted twice for each cycle and country – first using the original home possessions scores, then using the new home possessions scores. The cross-country comparability of the home possessions scores as a measure of SES was assessed with the variability of countries' component loadings for home possessions on SES, with a lower standard deviation indicating a higher level of cross-country comparability of the home possessions scores as a measure of SES.

Research question 4: Are the new home possessions scores a better predictor of students' cognitive scores on PISA than the original home possessions scores? This research question was addressed in Study 4. To assess whether the new home possessions scores were a better predictor of students' cognitive scores on PISA than the original home possessions scores, students' scores on the PISA reading, math, and science assessments were predicted separately by the original and new home possessions scores. If a larger percentage of variation in students' scores was explained by the new home possessions scores than the original home possessions scores, this was taken as evidence that the new home possessions scores were a more accurate measure of SES, compared to the original home possessions scores.

Research question 5: What evidence can be collected to support the external validity of the new home possessions scores as a measure of SES? This research question was addressed in Study 5. To collect validity evidence supporting the use of the

new home possessions scores as a measure of SES, for each cycle, the average of each country's new home possessions scores was correlated with the country's Human Development Index (HDI), a composite index developed by the United Nations Development Programme (UNDP) to measure different aspects of a country's development level ("Human Development Index," n.d.). Since HDI and SES are measured using similar components, a high correlation between the new home possessions scores and HDI was taken as evidence supporting the external validity of the new home possessions scores as a measure of SES.

CHAPTER 2 – METHODS

Data

Data for this research were drawn from publicly available datasets of PISA. While countries' participation in PISA is voluntary, the number of participating countries has steadily increased over the years. In the first cycle of PISA administered in 2000, 29 OECD and 14 non-OECD countries participated ("PISA 2000," n.d.). These numbers increased to 34 OECD and 38 non-OECD countries in the sixth cycle of PISA administered in 2015, the last cycle for which data are currently available ("PISA 2015," n.d.).[2]

While the current study tried to include as many countries as possible in the analyses, there were some criteria for exclusion. In 2000 and 2003, countries with a sizeable minority language population were excluded because there was no information in the public dataset on the language of examination for these cycles, making it impossible to divide these countries into country-by-language groups for Study 2.[3] As a result, 10 countries (out of 43 countries) were excluded from the final dataset in 2000,[4] while nine countries (out of 41 countries) were excluded from the final dataset in 2003.[5]

---

[2] In the seventh cycle of PISA administered in 2018, 37 OECD and 42 non-OECD countries participated. However, data from 2018 were not included in the current study because they were not publicly available at the time the study was conducted.

[3] In this research, a country with a sizeable minority language population is defined as a country in which there was at least one minority language that was used as the language of examination by at least 5% of the test takers of the country (using final student weights), based on PISA data from 2006 to 2015.

[4] Countries that were excluded in 2000 are Belgium, Canada, Finland, Israel, Latvia, Luxembourg, Macedonia, Romania, Spain, and Switzerland.

[5] Countries that were excluded in 2003 are Belgium, Canada, Finland, Latvia, Luxembourg, Slovak Republic, Spain, Switzerland, and Yugoslavia.

In addition, countries that had been excluded from the public dataset due to data adjudication issues, political issues, or other issues were also naturally excluded from the final dataset. Lastly, data from samples that were not nationally representative, such as data from specific regions or cities within a country, were excluded from the final dataset.[6] Appendix A lists the 75 countries that were included in the final dataset as well as the unweighted sample size for each country.

The target population for PISA was 15-year-old students attending educational institutions in grades 7 and higher, including foreign students, students attending foreign schools in the country, students enrolled on a part-time basis, and students attending vocational training programs and other related types of educational programs (OECD, 2017, p. 66).[7] Within each country, the selected students were weighted so the sample would be nationally representative. These weights were later readjusted so that within each cycle, the sum of the student weights for each country would be equal, regardless of the size of the target population in each country. In other words, within each cycle, all

[6] Regions or cities that participated in PISA include Beijing, Guangdong, Hong Kong, Jiangsu, Macao, and Shanghai in China; Himachal Pradesh and Tamil Nadu in India; Perm in Russia; regions in Spain; Massachusetts and North Carolina in the United States of America; and Miranda in Venezuela.

[7] In 2015, the national project manager of each country was responsible for constructing the school sampling frame that corresponded to the target population. While some schools and students were allowed to be excluded from the target population, the overall exclusion rate within a country was not allowed to exceed 5% of the desired target population. For the sampling of students, two-stage stratified sampling was used in all countries except Russia. In the first stage, schools were sampled from the school sampling frame, with the probability of selection proportional to the size of the school. Within a country, if less than 85% of the selected schools agreed to participate in the assessment, replacement schools were selected. After replacement, the school participation rate was required to be at least 65% for each country. In the second stage of sampling, students were sampled from the selected schools, with equal probability of selection for all students. Within a school, at least 50% of the selected students had to participate in the assessment in order for the school to be considered a participating school. Among the participating schools, the overall student response rate was required to be at least 80%. More information on the sampling methodology and results can be found in Chapter 4 and Chapter 11 of the PISA 2015 Technical Report (OECD, 2017).

countries were weighted equally, in line with the method that was used to scale the PISA cognitive scores (OECD, 2017, p. 291). Subsequently, the student weights were adjusted again so that the sum of the student weights for each cycle would be equal, regardless of the number of countries that participated in each cycle.[8] This weighting method ensured that each cycle would contribute equally to the analyses, while in each cycle, each country would contribute equally.

Measures

In PISA, students' SES was measured with information they provided on their home possessions (representing family wealth), parents' education, and parents' occupation (OECD, 2017, p. 339). This information was collected through the student questionnaire which was administered directly to students after the cognitive assessments (OECD, 2017, p. 36). The student questionnaire was designed to take no longer than 35 minutes, with 30 minutes allocated to the international questionnaire and an additional five minutes for any country-specific questions.

Home Possessions. To measure home possessions, students were asked whether they possessed or had access to certain items at home. Some items were dichotomous (i.e., it asked whether the student's household possessed the item or not), while other items had polytomous ordinal responses (i.e., it asked how many of the item the student's

[8] Specifically, the student weights were readjusted so the sum would be 5,000,000 for each cycle. Since 32 countries participated in PISA 2000, the student weights for this cycle were readjusted so the sum of the student weights for each country that participated in this cycle would be 156,250 (which is 5,000,000 divided by 32). For PISA 2015, since 65 countries participated in this cycle, the student weights for this cycle were readjusted so the sum of the student weights for each country that participated in this cycle would be 76,923 (which is 5,000,000 divided by 65).

household possessed). In almost every cycle of PISA, a few items were added to or dropped from the home possessions scale, taking into account the social, economic, and technical changes in the participating countries (OECD, 2017, p. 341). Also, each country was allowed to include up to three country-specific items in the scale.[9]

In this study, some items in the scale were excluded or modified to maintain the comparability of the scale across cycles and countries. For example, in 2000, the item regarding a calculator was a polytomous item, but it was recoded as a dichotomous item to make it consistent with the other cycles. Also, in 2000, the response categories for the number of books at home were different from the other cycles, and it was impossible to recode the responses to make them consistent with the other cycles. Therefore, this item was excluded from the final dataset for 2000. In 2003, TV, car, bathroom, cell phone, and computer (polytomous) were in the student questionnaire, but the responses were not included in the public dataset. This was also the case for bathroom in 2006. Consequently, these items were not present in the final dataset for the respective cycles. In 2006, an item asked whether students had a DVD or VCR player at home, while the item in the latter cycles only asked about DVD players. Since it was impossible to determine whether each student had a DVD player at home in 2006, this item was excluded from the final dataset for 2006. Lastly, all country-specific items were excluded from the final dataset for all cycles because these items were not comparable across countries. Table 1 presents the items that were included in the final dataset in each cycle.

[9] Examples of country-specific items in PISA 2015 include a guest room in the United States of America, solar panels in Australia, a jacuzzi in Russia, an espresso machine in Israel, a refrigerator with a freezer in Uruguay, and an air conditioner in Vietnam (OECD, 2017, p. 436).

Table 1

Items Included in the Home Possessions Scale in the Final Dataset

	Cycle						# of cycles
	2000	2003	2006	2009	2012	2015	
Desk to study at	✓	✓	✓	✓	✓	✓	6
Room of your own	✓	✓	✓	✓	✓	✓	6
Quiet place to study	✓	✓	✓	✓	✓	✓	6
Computer you can use for school work		✓	✓	✓	✓	✓	5
Educational software	✓	✓	✓	✓	✓	✓	6
Link to the internet	✓	✓	✓	✓	✓	✓	6
Classic literature	✓	✓	✓	✓	✓	✓	6
Books of poetry	✓	✓	✓	✓	✓	✓	6
Works of art	✓	✓	✓	✓	✓	✓	6
Books to help with your school work	✓ [a]	✓	✓	✓	✓	✓	6
Technical reference books				✓	✓	✓	3
Dictionary	✓	✓	✓	✓	✓	✓	6
Books on art, music or design						✓	1
Your own calculator	✓ [b]	✓	✓				3
Dishwasher	✓	✓	✓	✓	✓		5
DVD player				✓	✓		2
Television *	✓		✓	✓	✓	✓	5
Car *	✓ [c]		✓	✓	✓	✓	5
Room with a bath or shower *	✓ [d]			✓	✓	✓	4
Cellular phone *	✓		✓	✓	✓	✓ [e]	5
Computer *	✓		✓	✓	✓	✓	5
Tablet computer *						✓	1
E-book reader *						✓	1
Musical instrument *	✓					✓	2
Books * [f]		✓	✓	✓	✓	✓	5

Note. Polytomous items are indicated with an asterisk. Except for the number of books (the last item), the response categories for the polytomous items were zero, one, two, and three or more.

[a] This item asked whether the student had a textbook at home. Although the wording was not consistent with the other cycles, it was included in the final dataset. [b] This was a polytomous item, but it was recoded as a dichotomous item in the final dataset to make it consistent with the other cycles. [c] This item asked how many motor cars the student had at home. Although the wording was not consistent with the other cycles, it was included in the final dataset. [d] This item asked how many bathrooms the student had at home. Although the wording was not consistent with the other cycles, it was included in the final dataset. [e] This item asked how many cellular phones with internet access the student had at home. Although the wording was not consistent with the other cycles, it was included in the final dataset. [f] The response categories for this item were zero to 10, 11 to 25, 26 to 100, 101 to 200, 201 to 500, and more than 500.

Missing data. The percent of the sample with data on each item of each cycle is presented in Appendix B. For all items and cycles, data were missing for 5% or less of the sample.

Scaling method. In all cycles, item response theory (IRT) was used to scale the items and to generate the home possessions score for each student, with the latent trait (θ) defined as family wealth. These scores were subsequently included in the PCA to generate each student's SES score (explained below).

It should be noted that the exact method to scale the home possessions items was not consistent throughout the cycles. In 2000, instead of generating a single score for home possessions, separate scores were generated for household wealth, cultural possessions, and home educational resources. Also, the item endorsement parameter (β) for each item was estimated on the combined OECD sample, using the Rasch model (Rasch, 1960) to scale the dichotomous items and the partial credit model (PCM; Masters, 1982) to scale the polytomous items (OECD, 2002). In 2003, a single score was generated for home possessions, again estimating the item endorsement parameters (β) on the combined OECD sample (OECD, 2005). In 2006, due to the high level of between-country variation in the item endorsement parameters (β), the parameters were estimated separately for each country, constraining the sum of the parameters in each country to zero (OECD, 2009). In 2009, the item endorsement parameters (β) were estimated within each country using data from all the cycles the country had participated in, with each cycle weighted equally. Subsequently, a linear transformation was applied to each country's parameter to place them on a common scale (OECD, 2012). In 2012,

the item endorsement parameters (β) were again estimated using data from all previous cycles, but the relative position of each country was estimated on a joint scale (OECD, 2014b). Lastly, in 2015, the two-parameter logistic (2PL) model (Birnbaum, 1968) was used to scale the dichotomous items, while the generalized partial credit model (GPCM; Muraki, 1992) was used to scale the polytomous items, allowing each item to have its own discrimination parameter (α) as well as its own endorsement parameter (β). These parameters were estimated using data only from the 2015 cycle. Also, to address DIF across countries, for each item, if a country had an observed ICC which exhibited substantial misfit with the model-based ICC (which was estimated with data from all countries), indicated by a root mean square deviation (RMSD) value of over 0.3, the country was allowed to estimate its own item discrimination parameter (α) and item endorsement parameter (β) for the item (OECD, 2017, p. 296).

Parents' Education. To measure parents' education, students were asked about the highest level of schooling that their mother and father had completed. The response choices were based on the International Standard Classification of Education (ISCED) framework established by UNESCO in 1997 which classifies educational qualifications into primary, lower secondary, vocational/prevocational upper secondary, general upper secondary, non-tertiary postsecondary, vocational tertiary, and theoretically oriented tertiary/postgraduate education ("ISCED 1997," 1997). Subsequently, the higher ISCED level of either parent was recoded into the estimated years of schooling, based on the education system of each country at the time of the survey (OECD, 2017, p. 298). This

score, representing the highest educational level of the parents, was subsequently included in the PCA to generate each student's SES score (explained below).

Parents' Occupation. To measure parents' occupation, students were asked open-ended questions about their mother and father's occupation. These responses were later mapped onto the International Standard Classification of Occupations (ISCO) framework established by the International Labour Organization (ILO). Subsequently, the occupations were converted into numeric scores using the International Socio-Economic Index of Occupational Status (ISEI) framework which assigns a higher score to occupations with a higher status (OECD, 2017, p. 298). The higher ISEI score of either parent, representing the highest occupational level of the parents, was subsequently included in the PCA to generate each student's SES score (explained below).

It should be noted that the ISCO and ISEI frameworks were updated in 2008. Thus, for the PISA cycles from 2000 to 2009, parents' occupational status was measured using the ISEI framework established in 1988, but from the 2012 cycle, it was measured using the framework established in 2008 (OECD, 2014b, p. 55).

SES scores. To generate students' SES scores, PCA was conducted with the three components mentioned above. It was assumed that the first principal component represented SES, so the component score on the first principal component was taken as each student's SES score, which is called the economic, social and cultural status (ESCS) score in PISA (OECD, 2017, p. 339). Within each cycle, a single component loading was

used to weight each component for all countries (OECD, 2012, p. 315),[10] even though the actual relationship between each component and the first principal component varied across countries.

It should be noted that the countries included in the PCA to estimate the component loadings were not consistent across the PISA cycles. Until 2012, the PCA only included OECD countries, and the SES scores were standardized by constraining the mean score to zero and the standard deviation to one for the OECD countries (OECD, 2014b, p. 352). For non-OECD countries, the SES scores were generated using the component loadings for each component (which had been estimated only with the OECD countries), the student's score for each component (which had been standardized on the OECD countries), and the country's eigenvalue for the first principal component. However, in 2015, the PCA to estimate the component loadings included all participating countries, but the SES scores were still standardized only on the OECD countries (OECD, 2017, p. 340).

Analyses

Study 1: Measurement invariance across cycles. The purpose of this study was to determine which items demonstrated measurement invariance across the six cycles of PISA. The software mdltm (version 1.965; von Davier, 2005) was used for this analysis because it has an algorithm that automatically assesses the partial invariance of the model

[10] For example, in 2009, a component loading of 0.74 was used for home possessions, 0.81 for parents' education, and 0.81 for parents' occupation for all countries that participated in this cycle (OECD, 2012, p. 315).

and assigns unique parameters to groups that exhibit misfit. This software was also used to scale the items for the cognitive assessment and background questionnaires for PISA 2015 (OECD, 2017, p. 144) and PISA 2018.

Estimating the total sample ICC. In this step, all the items in the home possessions scale were calibrated concurrently using data from all cycles and all countries. As explained above, each cycle was weighted equally, and within each cycle, all the countries were weighted equally. Missing data, whether it was because an item was not administered in a cycle, a country did not participate in a cycle, or a student did not respond to an item, were treated as ignorable missing data (Shin, Khorramdel, Xu, & von Davier, 2017).

Dichotomous items were scaled using the 2PL model which assumes that the probability that a subject (s) owns an item (i) at a given level of a latent trait (θ) depends on the item's discrimination (α) and endorsement (β) parameters, as expressed in the following equation (Embretson & Reise, 2000):

$$P(X_{is} = 1 \mid \theta_s, \beta_i, \alpha_i) = \frac{e^{\alpha_i * 1.7(\theta_s - \beta_i)}}{1 + e^{\alpha_i * 1.7(\theta_s - \beta_i)}} \quad (1)$$

Polytomous items were scaled using the GPCM which assumes that the probability that a subject selects a response category (X) of an item (i) at a given level of a latent trait (θ) depends on the item's discrimination parameter (α), the item's step

endorsement parameters (δ_{ij}),[11] and information from all the other response categories, as expressed in the following equation (Embretson & Reise, 2000):

$$P_{ix}(\theta) = \frac{e^{\sum_{j=0}^{x} \alpha_i * 1.7(\theta - \delta_{ij})}}{\sum_{r=0}^{M}[e^{\sum_{j=0}^{r} \alpha_i * 1.7(\theta - \delta_{ij})}]} \quad (2)$$

In the first round of the item calibration, for each item, the item parameters were estimated with data pooled from all cycles and all countries, producing a model-based ICC, as illustrated in Figure 1. In this research, this model-based ICC will be called the total sample ICC, since it was estimated with data from the entire sample.

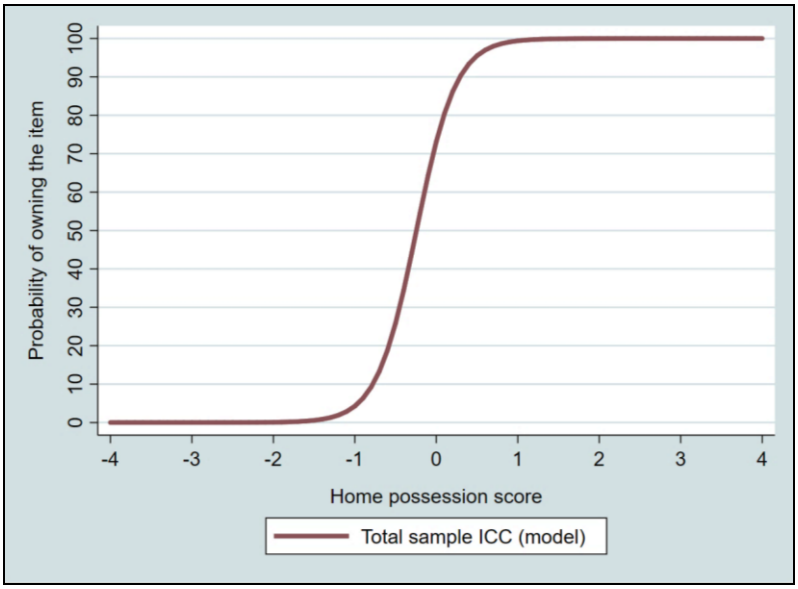


Figure 1. The total sample ICC estimated with data from the entire dataset. This figure is for illustrative purposes only.

[11] An item's step endorsement parameter (δ_{ij}) is the intersection between a response category and an adjacent response category.

All the items were calibrated concurrently, placing them on a common scale measuring the latent trait (θ), family wealth. To solve the indeterminacy of the IRT scale, the average of the item discrimination parameters (α) across the items was constrained to one, while the average of all the intercepts across the items was constrained to zero.[12]

***Detecting DIF across cycles and assigning unique item parameters*.** DIF across cycles was detected using multiple group concurrent calibration with partial invariance constraints which detects DIF in an IRT framework, in line with the method used in PISA 2015 (OECD, 2017, p. 143). In this study, groups were defined as the cycles, so data from all countries that participated in a cycle were pooled together to form a group. In Figure 2, the total sample ICC is represented by the red curve, while the observed ICCs for the cycles are represented by orange curves, with darker curves representing ICCs for more recent cycles.[13]

[12] Dichotomous items only have one intercept for which the value is -1.7 * α * β. The number of intercepts for polytomous items is one less than the number of response categories for the item, and the values are -1.7 * α * δ_{ij}.

[13] The observed ICCs are based on the pseudo counts from the E-step in the EM algorithm (Shin et al., 2017).

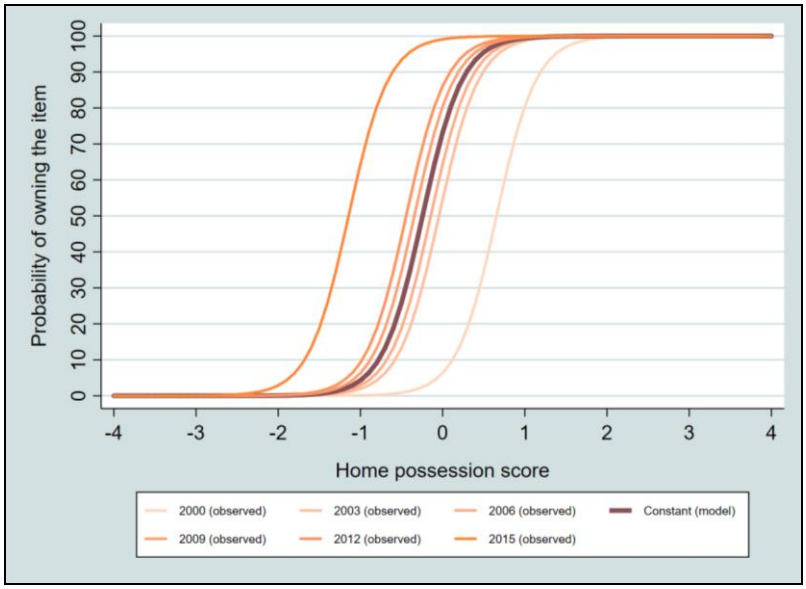


Figure 2. The total sample ICC estimated with data from all cycles and the observed ICC for each cycle. This figure is for illustrative purposes only.

To detect DIF across cycles, for each cycle, RMSD was estimated using the following equation which quantifies the difference between the model-based ICC ($P_e(\theta)$) and the observed ICC for each cycle ($P_o(\theta)$), weighted by the θ distribution (Shin et al., 2017):[14]

$$\text{RMSD} = \sqrt{\int (P_o(\theta) - P_e(\theta))^2 \, f(\theta) d\theta} \quad (3)$$

RMSD values are always positive or zero because it indicates the absolute difference between the model-based ICC and the observed ICC for each group (i.e., cycle

[14] In the first round of item calibration, the model-based ICC was the total sample ICC. From the second round of item calibration, the model-based ICC was either the total sample ICC, the semi-total sample ICC, or the cycle-specific ICC, depending on whether the cycle (for which the RMSD was being estimated) had been assigned unique item parameters in the previous round(s) of item calibration.

in this study), and a higher value of RMSD indicates a higher level of misfit between the model-based ICC and the observed ICC.

In the first round of the item calibration, a cycle with an RMSD value greater than 0.40 was considered to exhibit substantial misfit with the model-based ICC. In other words, DIF was detected for the cycle, so measurement invariance could not be established across the cycles for the item. In the subsequent rounds of item calibration, cycles which had exhibited DIF were assigned unique item discrimination (α) and item endorsement parameters (β) estimated with data only from the respective cycle, resulting in item parameters that better fit the cycle's observed ICC. In this research, the ICCs for these cycles will be called unique ICCs or cycle-specific ICCs, since the item parameters were estimated specifically for those cycles. For the remaining cycles, the ICC was estimated again, pooling data across the remaining cycles. This ICC will be called the semi-total sample ICC, since it was estimated with data from most, but not all, of the cycles.

The process of assigning unique item parameters to cycles that exhibited DIF was repeated using RMSD cutoff values of 0.35, 0.30, 0.25, 0.20, and 0.15 until all cycles had RMSD values below 0.15.

Subsequently, the mean deviations (MD) were also examined to detect any remaining DIF across cycles. MD is similar to RMSD in that it quantifies the difference between the model-based ICC ($P_e(\theta)$) and the observed ICC for each cycle ($P_o(\theta)$), weighted by the θ distribution. However, MD also takes into account the direction of the deviation, using the following equation (Shin et al., 2017):

$$MD = \int (P_o(\theta) - P_e(\theta)) f(\theta) d\theta \quad (4)$$

Unlike RMSD, MD values can be either positive or negative, with values further from zero indicating larger misfit between the model-based ICC and the observed ICC. Positive MD values indicate that in general, the observed ICC lies above the model-based ICC (i.e., a higher proportion of subjects owned the item than what was predicted by the model), while negative MD values indicate that in general, the observed ICC lies below the model-based ICC (i.e., a lower proportion of subjects owned the item than what was predicted by the model). Compared to RMSD, MD is less sensitive to differences in the slopes of the model-based and observed ICCs (Shin et al., 2017). This is because when the model-based ICC and the observed ICC cross each other due to a difference in the slopes, the MD will be positive on one side of the point at which the curves cross, while negative on the other side. The positive and negative MD values will cancel each other out, resulting in an overall MD value that is closer to zero. Nevertheless, MD provides valuable information in that it indicates the direction of the misfit.

When all cycles have RMSD and MD values below 0.15, the model-based ICC for each cycle will adequately fit its observed ICC, as illustrated in Figure 3. In this figure, each red curve represents the model-based ICC for each cycle(s), with the lightest curve representing the model-based ICC for 2000, the darkest curve representing the model-based ICC for 2015, and the middle curve representing the model-based ICC for the remaining cycles. This is essentially a partial invariance model (Byrne et al., 1989) in

which most cycles are constrained to have the same item parameters, while certain cycles are assigned unique item parameters. These results made it possible to analyze whether and how the item parameters shifted across the cycles, since the item parameters for all cycles were estimated on a common scale.

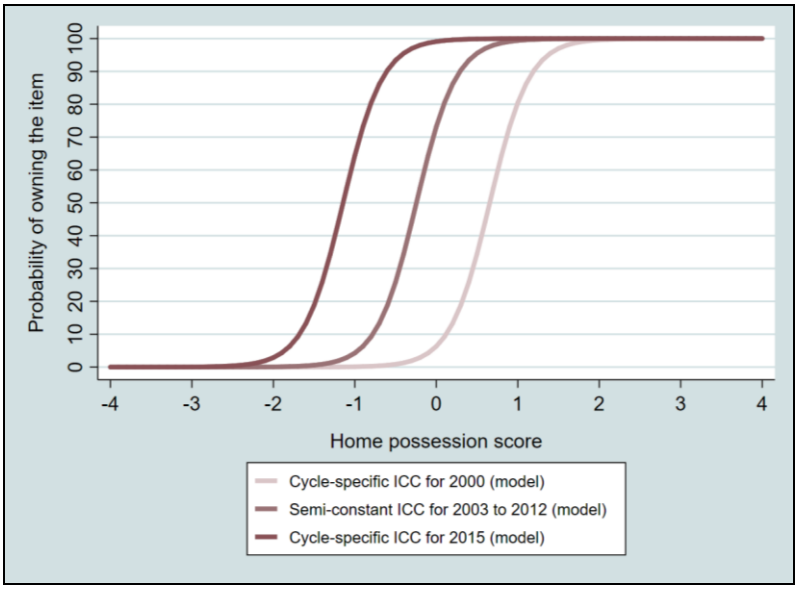


Figure 3. The cycle-specific ICCs for cycles exhibiting DIF and the semi-total sample ICC for the remaining cycles. This figure is for illustrative purposes only.

It should be noted that assigning unique item parameters to more cycles will increase the accuracy of the home possessions scores within each cycle because the item parameters for each cycle will more accurately represent the true relationship between the possession of an item and family wealth for the cycle. However, when more cycles are assigned unique item parameters, the cross-cycle comparability of the home possessions scores as a measure of family wealth will decrease because the number of common parameters used across the cycles to measure the latent variable will decrease. Therefore,

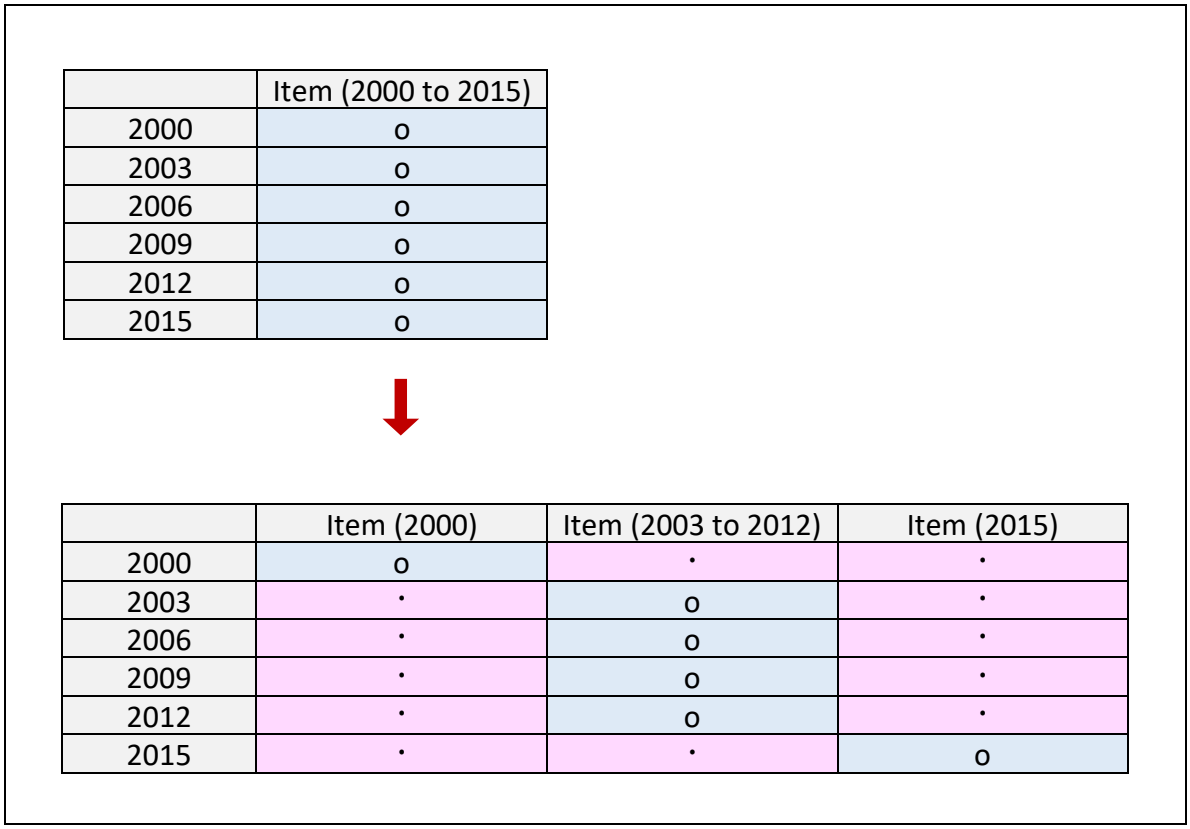
there is a trade-off between the accuracy of scores within each cycle and the comparability of scores across the cycles (OECD, 2017, p. 227).

Study 2: Measurement invariance across country-by-language groups. The purpose of Study 2 was to determine which items demonstrated measurement invariance across the country-by-language groups. To take into account DIF across cycles, the final model of Study 1 was used as the initial model of Study 2. Again, the software mdltm (version 1.965; von Davier, 2005) was used for the analysis.

For this study, students within countries were grouped by the language in which they took the cognitive assessment because it was hypothesized that the relationship between the possession of an item and family wealth depended on sociocultural factors which were partially captured by the language of examination. Languages that were used as the language of examination by at least 5% of the test takers in the country (using final student weights) were considered to be independent country-by-language groups, while languages that were used as the language of examination by less than 5% of the test takers were combined with the majority language group of the country. This created a total of 96 country-by-language groups, as presented in Appendix C. No adjustments were made to the student weights after students were grouped into country-by-language groups.

To use the final model of Study 1 as the initial model, the dataset had to be reshaped. Figure 4 illustrates how the dataset was reshaped for an illustrative item which exhibited DIF in 2000 and 2015. For this item, a separate column was inserted for 2000 and 2015, and data from 2000 and 2015 were copied into their respective columns. In the

column which had originally contained data for this item, the cells for 2000 and 2015 were left missing. As stated above, missing data were treated as ignorable missing data during the item calibration process.



	Item (2000 to 2015)
2000	o
2003	o
2006	o
2009	o
2012	o
2015	o

	Item (2000)	Item (2003 to 2012)	Item (2015)
2000	o	·	·
2003	·	o	·
2006	·	o	·
2009	·	o	·
2012	·	o	·
2015	·	·	o

Figure 4. Reshaping the dataset for an illustrative item which exhibited DIF in 2000 and 2015. o = Non-missing data. · = Missing data.

Estimating the international ICC. As in Study 1, dichotomous items were scaled using the 2PL model, and polytomous items were scaled using the GPCM. In the first round of item calibration, the items were calibrated using data pooled across all countries (i.e., none of the country-by-language groups were assigned unique item parameters). Nevertheless, due to the shape of the dataset, a cycle which had exhibited DIF in Study 1

were assigned unique item parameters (which were estimated with data pooled across all countries that participated in the cycle). As a result, the model-based ICCs produced in this round of item calibration were similar to the model-based ICCs produced in the final model of Study 1, as illustrated in Figure 5.[15] In this research, these ICCs will be called international ICCs, since they were estimated with data from all country-by-language groups.

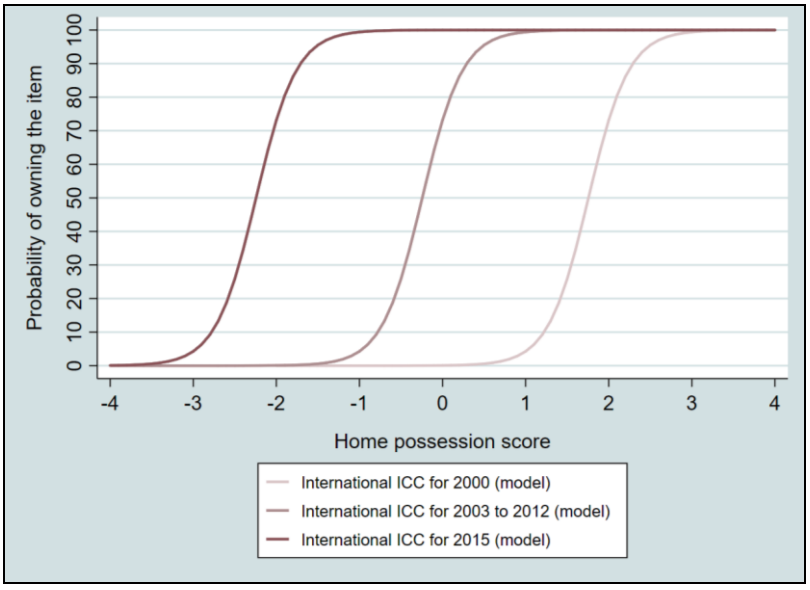


Figure 5. The international ICCs estimated with data from all country-by-language groups. This figure is for illustrative purposes only.

Detecting DIF across country-by-language groups and assigning unique item parameters. As in Study 1, DIF was detected using multiple group concurrent calibration with partial invariance constraints, but in this study, the groups were defined as the

[15] The item parameters of these two models are not identical, due to the different number of columns in the two datasets. In mdltm, each column is considered to be a separate item, and the average of the item discrimination parameters (α) are constrained to one, while the average of all the intercepts are constrained to zero.

country-by-language groups. In Figure 6, the international ICCs are represented by the red curves, while the observed ICCs for each country-by-language group are represented by the green curves, with darker curves representing ICCs for more recent cycles.

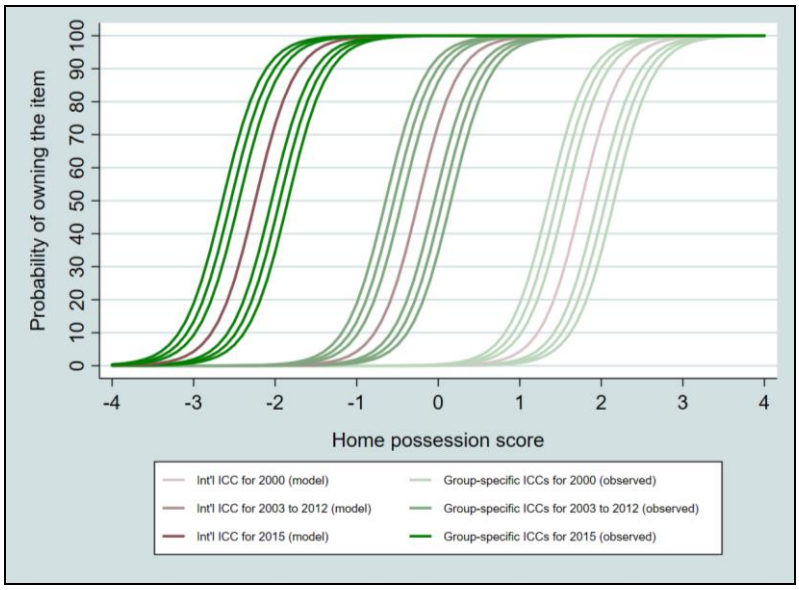


Figure 6. The international ICCs estimated with data from all country-by-language groups and the observed ICCs for each country-by-language group. This figure is for illustrative purposes only.

To detect DIF across the country-by-language groups, the observed ICC for each country-by-language group was compared to the model-based ICC.[16] In the first round of item calibration, a country-by-language group with an RMSD value greater than 0.40 was considered to exhibit substantial misfit with the model-based ICC. In other words, DIF was detected for the country-by-language group, so measurement invariance could not be established across the country-by-language groups for the item. In the subsequent rounds

[16] In the first round of item calibration, the model-based ICC was the international ICC. From the second round of item calibration, the model-based ICC was either the international ICC, the semi-international ICC, or the group-specific ICC, depending on whether the country-by-language group (for which the RMSD was being estimated) had been assigned unique item parameters in the previous round(s) of item calibration.

of item calibration, country-by-language groups which had exhibited DIF were assigned unique item discrimination (α) and item endorsement parameters (β) estimated with data only from the respective country-by-language group, resulting in item parameters that better fit the country-by-language group's observed ICC. In this research, the ICCs for these country-by-language groups will be called unique ICCs or group-specific ICCs, since the parameters were estimated specifically for those country-by-language groups. For the remaining country-by-language groups, the ICC was estimated again, pooling data across the remaining country-by-language groups. This ICC will be called the semi-international ICC, since it was estimated with data from most, but not all, of the country-by-language groups.

The process of assigning unique item parameters to country-by-language groups that exhibited DIF was repeated using RMSD cutoff values of 0.35, 0.30, 0.25, 0.20, and 0.15, as well as an MD cutoff value of 0.15. At the end of this process, all the country-by-language groups had RMSD and MD values below 0.15, indicating that the model-based ICC for each country-by-language group adequately fit its observed ICC, as illustrated in Figure 7. In this figure, the red curves represent the semi-international ICCs and the blue curves represent the group-specific ICCs, with darker curves representing ICCs for more recent cycles. This made it possible to analyze whether and how the item parameters varied across the country-by-language groups, since the item parameters for all the country-by-language groups were estimated on a common scale.

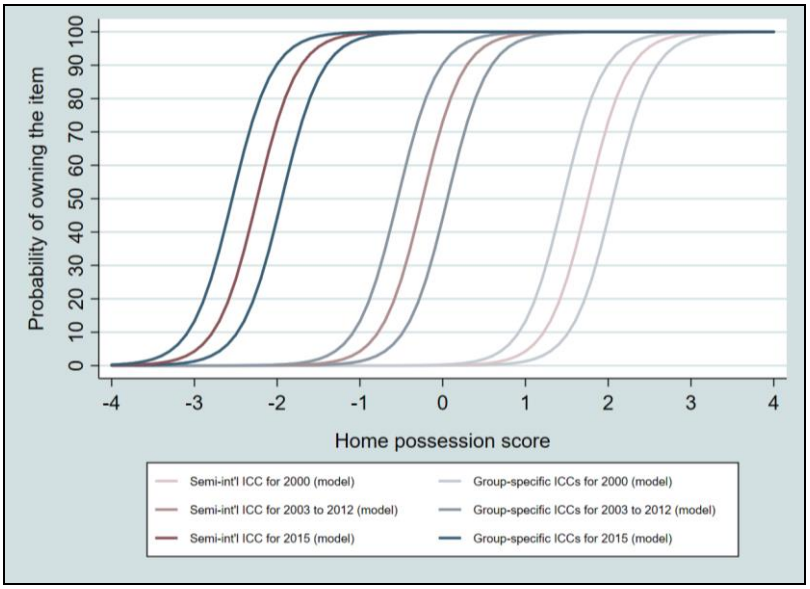


Figure 7. The group-specific ICCs for country-by-language groups exhibiting DIF and the semi-international ICC for the remaining country-by-language groups. This figure is for illustrative purposes only.

Again, it should be noted that assigning unique item parameters to more country-by-language groups will increase the accuracy of the home possessions scores within each country-by-language group because the item parameters for each country-by-language group will more accurately represent the true relationship between the possession of an item and family wealth for the country-by-language group. Table 2 summarizes the effect of different methods on the accuracy of the home possessions scores within country-by-language groups.

Table 2

Effect of Different Methods on the Accuracy of the Home Possessions Scores within Country-by-Language Groups

Method (Cycles in which the method was used only for the new scale)	Accuracy of scores within country-by-language groups	
	Decrease	Increase
Constrain item parameters to be equal across the country-by-language groups by default (2006 to 2012) *	✓	
Calibrate item parameters with data from all years (2006 and 2015)	✓	
Use 2PL model and GPCM to calibrate the items (2006 to 2012)		✓
Use a lower cutoff to detect DIF (2015) *		✓

Note. Methods that also affect the comparability of scores across country-by-language groups are indicated with an asterisk.

However, when more country-by-language groups are assigned unique item parameters, the comparability of the home possessions scores as a measure of family wealth across the country-by-language groups will decrease because the number of common item parameters used across the country-by-language groups to measure the latent variable will decrease. Thus, as mentioned above, there is a trade-off between the accuracy of scores within the country-by-language groups and the comparability of scores across the country-by-language groups (OECD, 2017, p. 226). Table 3 summarizes the effect of different methods on the comparability of the home possessions scores across the country-by-language groups.

Table 3

Effect of Different Methods on the Comparability of the Home Possessions Scores across Country-by-Language Groups

Method (Cycles in which the method was used only for the new scale)	Comparability of scores across country-by-language groups	
	Decrease	Increase
Constrain item parameters to be equal across the country-by-language groups by default (2006 to 2012)		✓
Use a lower cutoff to detect DIF (2015)	✓	

Study 3: Cross-country comparability of the new home possessions scores as a measure of SES. The purpose of this study was to assess whether the new home possessions scores (generated from the final model in Study 2) were a more comparable measure of SES across countries than the original home possessions scores (obtained from the public dataset). To assess this, for each cycle and country, PCAs were conducted twice – first using the original home possessions scores, then using the new home possessions scores. The other components included in the PCAs were parents' education and parents' occupation.

The first principal component was assumed to represent SES. High variation across countries in the component loadings of home possessions on SES, measured by the standard deviation of the component loadings of home possessions on SES, was an indication that the relationship between home possessions and SES varied across countries, implying that the home possessions scores were not a comparable measure of

SES across countries. Conversely, a lower standard deviation of the component loadings of home possessions on SES was an indication that the home possessions scores were a more comparable measure of SES across countries. Thus, to assess whether the new home possessions scores were a more comparable measure of SES across countries than the original home possessions scores, the standard deviation of the component loadings of home possessions on SES when the new home possessions scores were used in the PCA was compared to the standard deviation of the component loadings of home possessions on SES when the original home possessions scores were used in the PCA. If the standard deviation was lower for the former, this was taken as evidence that the new home possessions scores were a more comparable measure of SES across countries than the original home possessions scores. Stata (version 15) was used for this study.

To allow for comparisons across cycles, only the countries that participated in all cycles of PISA from 2006 to 2015 were included in the study. Data from 2000 and 2003 were not used because it would have reduced the number of countries included in the study to 26. Also, missing data were not imputed for any of the components because the purpose of the analysis was to compare the results of the PCA using the original home possessions scores against the results of the PCA using the new home possessions scores. Using the same imputation method for both datasets may have imputed similar values for both datasets, decreasing the observed differences between the results of the two PCAs. As a consequence of not imputing data for any of the components, only the students with data on all three components of SES were included in the analysis. The 50 countries that were included in the analysis as well as the percent of the sample in each country that had

data on all three components of SES are presented in Appendix D. In 2006, an average of 7% of the sample were excluded from the analysis in each country (with a maximum of 42% in Qatar); in 2009, an average of 7% of the sample were excluded from the analysis in each country (with a maximum of 20% in Qatar); in 2012, an average of 8% of the sample were excluded from the analysis in each country (with a maximum of 24% in Germany); and in 2015, an average of 11% of the sample were excluded from the analysis in each country (with a maximum of 23% in Thailand).

It was hypothesized that for the cycles from 2006 to 2012, the new home possessions scores would be a more comparable measure of SES across countries than the original home possessions scores. This is because in Study 2 (in which the new home possessions scores were generated), the default was to constrain the item parameters to be equal across the country-by-language groups, and only the country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC were assigned unique item parameters. This is in contrast to the original method used to scale the items for these cycles, which estimated item parameters separately for each country. Thus, it was hypothesized that the new home possessions scale would be more comparable across countries than the original home possessions scale, and in the same logic, that the new home possessions scores would be a more comparable measure of SES across countries than the original home possessions scores.

For the 2015 cycle, it was hypothesized that the new home possessions scores would be a less comparable measure of SES across countries than the original home possessions scores. While the item parameters for both models were estimated using

similar methods (i.e., the 2PL model and the GPCM were used to calibrate the items, the item parameters were constrained to be equal across the country-by-language groups by default, and only the country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC were assigned unique item parameters), the cutoff used for detecting DIF was lower for the new model.[17] As a result, more country-by-language groups were assigned unique item parameters in the new model. Thus, it was hypothesized that the new home possessions scale would be less comparable across countries than the original home possessions scale, and in the same logic, that the new home possessions scores would be a less comparable measure of SES across countries than the original home possessions scores.

Study 4: Predicting PISA cognitive scores with the original and new home possessions scores. The purpose of this study was to assess whether the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores. To assess this, for each cycle and country, bivariate linear regressions were run twice to predict students' PISA scores in reading, math, and science – first using the original home possessions scores, then using the new home possessions scores. Since many studies have found that SES was a statistically significant predictor of students' academic achievement (Sirin, 2005), if the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores, this was taken as evidence that the new home possessions scores were a more accurate

[17] In the new model, country-by-language groups with an RMSD or MD value of 0.15 or above were considered to exhibit DIF, while in the original model, the cutoff was an RMSD value of 0.30 (OECD, 2017, p. 296).

measure of SES than the original home possessions scores. The International Database Analyzer (IDB Analyzer; version 4.0.23) developed by the IEA was used for the analysis, taking into account the student sampling weights and plausible values for the cognitive scores.[18]

For reasons explained above, only the 50 countries that participated in all cycles of PISA from 2006 to 2015 were included in the analysis, and missing data were not imputed. The countries that were included in the analysis as well as the percent of the sample in each country that had data on the home possessions scores are presented in Appendix E. For the cycles from 2006 to 2012, home possessions scores were missing for an average of 1% of the sample in each country (with a maximum of 4% in Qatar in 2006, 7% in Germany in 2009, and 15% in Germany in 2012), while in 2015, home possessions scores were missing for an average of 2% of the sample in each country (with a maximum of 13% in Germany).

It was hypothesized that for the cycles from 2006 to 2012, the predictive power of the new home possessions scores would be higher than the original home possessions scores. Even though constraining the item parameters to be equal across the country-by-language groups by default may have decreased the accuracy of the new home possessions scores within each country, using the 2PL model and the GPCM to scale the items (instead of the Rasch model and the PCM, respectively) may have greatly increased the accuracy of the new home possessions scores within each country. This is because these models estimate an item discrimination parameter (α) in addition to the item

[18] The IDB Analyzer can be downloaded from this website: https://www.iea.nl/data

endorsement parameter (β) for each item, resulting in a more accurate model of the relationship between the possession of an item and family wealth. Thus, it was hypothesized that the new home possessions scores would be a more accurate measure of family wealth than the original home possessions scores, resulting in a higher predictive power for the new home possessions scores when used to predict the PISA cognitive scores.

For the 2015 cycle, it was hypothesized that the predictive power of the new home possessions scores and the original home possessions scores would be similar. While the two scales shared some similarities (i.e., the item parameters were constrained to be equal across the country-by-language groups by default, a country-by-language group was assigned unique item parameters only if the observed ICC for the group exhibited substantial misfit with the international ICC, and the 2PL model and the GPCM were used to scale the items), using a lower cutoff to detect DIF in the new scale resulted in more country-by-language groups being assigned unique item parameters, increasing the accuracy of the scores for these groups. However, in the new scale, the item parameters were constrained to be equal across all cycles by default, and the 2015 cycle was not assigned unique item parameters unless the observed ICC for 2015 exhibited substantial misfit with the total sample ICC. This may have resulted in less accurate item parameters for the new scale compared to the original scale which estimated all item parameters specifically for 2015. Thus, it was hypothesized that the accuracy of the new home possessions scores and the original home possessions scores would be similar,

resulting in similar levels of predictive power when using either the original or new home possessions to predict the PISA cognitive scores.

Study 5: Evidence supporting the external validity of the new home possessions scores as a measure of SES. The purpose of this study was to collect evidence supporting the external validity of the new home possessions scores as a measure of SES. For this analysis, the average of the new home possessions scores was calculated for each country and cycle. Then, for each cycle, the countries' average new home possessions score was correlated with the country's HDI for that year.[19] As mentioned above, HDI is a composite index developed by the UNDP to measure different aspects of a country's development level ("Human Development Index," n.d.), and it is calculated by taking the geometric mean of three components – the logarithm of the Gross National Income (GNI) per capita,[20] representing a decent standard of living; the expected years of schooling (for school-age children) or the mean years of schooling (for adults aged 25 years or more), representing education; and the life expectancy at birth, representing a long and healthy life.[21] Since two of the three components used to measure HDI and SES are similar (i.e., income and education), a strong correlation between the average new home possessions scores and HDI was taken as evidence supporting the external validity of the new home possessions scores as a measure of SES.

For reasons explained above, only the countries that participated in all cycles of PISA from 2006 to 2015 were included in the analysis, and missing data were not

[19] HDI data were downloaded from this website: http://hdr.undp.org/en/data#

[20] The logarithm of GNI per capita was used because it reflected the diminishing marginal utility of income in improving human development ("Human Development Index," n.d.).

[21] To calculate the geometric mean of three values, the values are multiplied, then the cube root is taken.

imputed. Taiwan was eventually excluded from the analysis because HDI for Taiwan was not available for any of the years included in the analysis.

CHAPTER 3 – RESULTS AND DISCUSSION

Study 1: Measurement Invariance across Cycles

Table 4 presents the item discrimination parameter (α), the item endorsement parameter (β), and the step endorsement parameters (δ_j) for each item of the home possessions scale at the end of Study 1.[22]

Table 4

Final Item Parameters for Study 1

Item (Cycle)	Item discrim-ination (α)	Item endorse-ment (β)	Step endorsement (δ_j)
Desk (2000 to 2015)	0.82	-1.57	
Own room (2000 to 2015)	0.64	-1.26	
Quiet study place (2000 to 2015)	0.63	-1.75	
Computer (2003 to 2015)	2.96	-0.48	
Ed software			
Ed software (2000)	1.53	0.30	
Ed software (2003 to 2015)	0.89	0.29	
Internet			
Internet (2000)	2.06	0.56	
Internet (2003 to 2012)	2.42	-0.24	
Internet (2015)	2.20	-0.74	
Classic literature (2000 to 2015)	0.35	0.00	
Poetry books (2000 to 2015)	0.28	-0.13	
Artwork (2000 to 2015)	0.60	-0.07	
School books (2000 to 2015)	0.44	-2.22	
Reference books (2009 to 2015)	0.65	-0.05	

[22] In Study 1, the item parameters were constrained to be equal across the cycles by default, and only the cycles for which the observed ICC exhibited substantial misfit with the total sample ICC (defined as RMSD or MD values over 0.15) were assigned unique item parameters. As a result, the model-based ICC for each cycle adequately fit its observed ICC.

Dictionary (2000 to 2015)	0.74	-2.27					
Books on culture (2015)	0.61	0.08					
Calculator (2000 to 2006)	0.93	-1.82					
Dishwasher (2000 to 2012)	0.85	0.15					
DVD player (2009, 2012)	0.92	-1.35					
TV (2000, 2006 to 2015) *	0.59		-3.09	-0.40	0.35		
Car (2000, 2006 to 2015) *	0.74		-0.52	0.58	1.44		
Bathroom (2000, 2009 to 2015) *	0.72		-1.57	0.89	1.68		
Cellphone *							
Cellphone (2000) *	0.68		-0.01	0.41	0.56		
Cellphone (2006 to 2012) *	0.67		-1.65	-0.77	-1.49		
Cellphone (2015) *	0.73		-1.19	-0.22	-0.88		
Computer *							
Computer (2000) *	2.00		0.07	1.02	1.38		
Computer (2006 to 2009) *	1.95		-0.54	0.61	1.02		
Computer (2012 to 2015) *	1.55		-0.72	0.23	0.63		
Tablet (2015) *	0.63		0.15	1.22	1.07		
Ebook reader (2015) *	0.48		2.51	2.52	1.51		
Instrument (2000, 2015) *	0.43		0.82	1.34	0.61		
Books (2003 to 2015) *	0.29		-0.48	-0.73	1.62	1.22	1.89

Note. Polytomous items are indicated with an asterisk.

Four items exhibited DIF across cycles – educational software, internet, cell phone, and computer (polytomous). To assess whether the results had been affected by the differences in the countries that participated in each cycle of PISA, the analysis was conducted again with only the 26 countries that had participated in all six cycles of PISA.[23] Again, only the four items mentioned above exhibited DIF across cycles.

For educational software, the item discrimination parameter (α) decreased in 2003 (i.e., it was 1.53 in 2000, and 0.89 from 2003 to 2015), while the item endorsement parameter (β) remained relatively stable (i.e., it was 0.30 in 2000, and 0.29 from 2003 to

[23] The 26 countries that participated in all six cycles of PISA are Australia, Austria, Brazil, Czech Republic, Denmark, France, Germany, Greece, Hungary, Iceland, Indonesia, Ireland, Italy, Japan, Korea (South), Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Russia, Sweden, Thailand, United Kingdom, and the United States of America.

2015), as presented in Figure 8. This indicates that the relationship between the possession of educational software and family wealth became weaker in 2003.

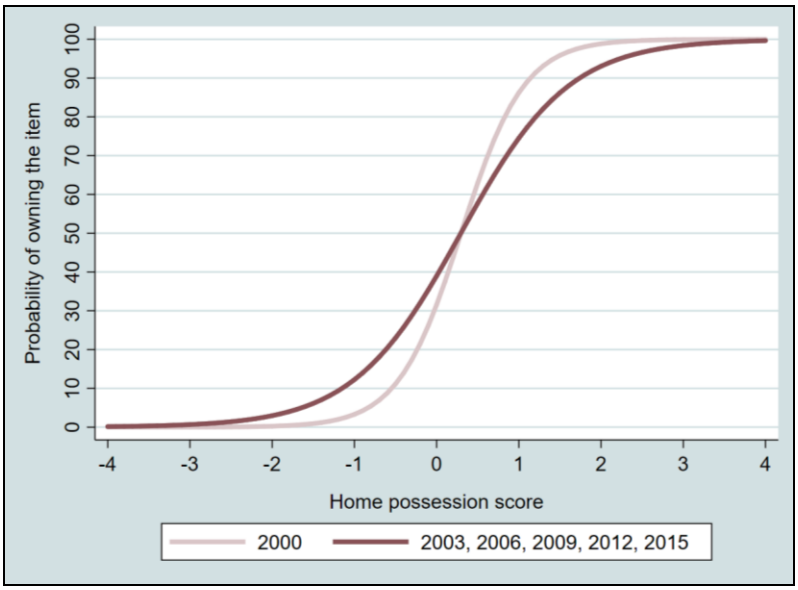


Figure 8. Model-based ICCs for educational software.

For internet access, the item discrimination parameter (α) remained relatively stable (i.e., it was 2.06 in 2000, 2.42 from 2003 to 2012, and 2.20 in 2015), but the item endorsement parameter (β) decreased in 2003 and 2015 (i.e., it was 0.56 in 2000, -0.24 from 2003 to 2012, and -0.74 in 2015), as presented in Figure 9. This indicates that internet became more accessible over the cycles.

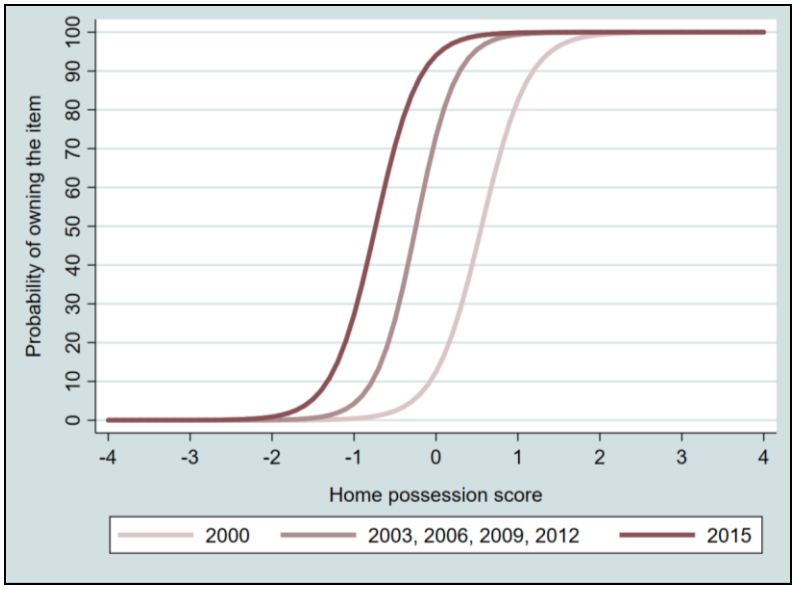


Figure 9. Model-based ICCs for internet.

For cell phone, the item discrimination parameter (α) remained relatively stable (i.e., it was 0.68 in 2000, 0.67 from 2006 to 2012, and 0.73 in 2015), while the item endorsement parameter (β) decreased in 2006 (i.e., it was 0.32 in 2000, and -1.31 from 2006 to 2012), indicating that cell phones became more accessible in 2006. However, in 2015, the item endorsement parameter (β) increased (i.e., it was -1.31 in 2006, and -0.76 in 2015), indicating that cell phones became less accessible in 2015. However, this may have been due to the fact that the item in 2015 asked about cell phones with internet access, while in the other cycles, the item did not mention internet access. Thus, the change in the wording of the item, not the relationship between the possession of cell phones and family wealth, may have caused the shift in the item endorsement parameter (β) in 2015.

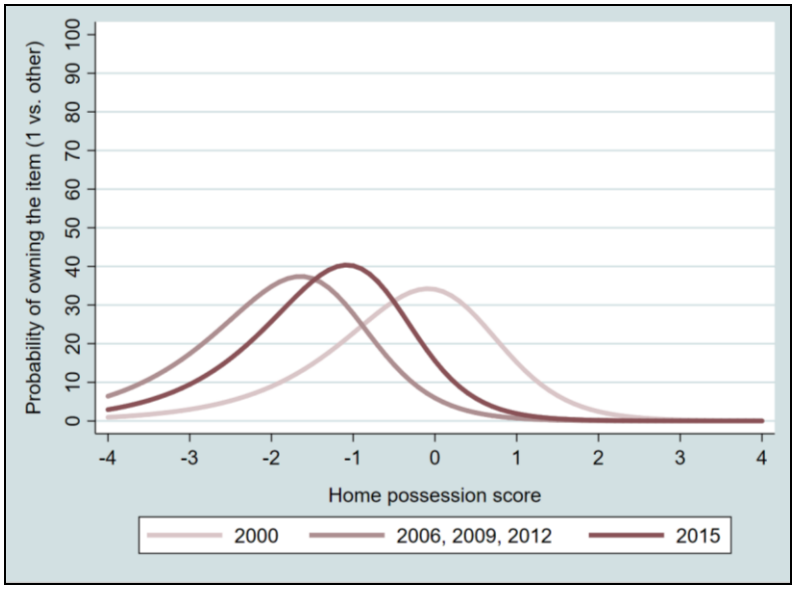


Figure 10. Model-based category response curves for cell phone. For simplicity, only one category response curve is shown for each cycle.

For computer (polytomous), the item discrimination parameter (α) decreased throughout the cycles (i.e., it was 2.00 in 2000, 1.95 from 2006 to 2009, and 1.55 from 2012 to 2015), while the item endorsement parameter (β) also decreased (i.e., it was 0.82 in 2000, 0.36 from 2006 to 2009, and 0.05 from 2012 to 2015), as presented in Figure 11. This indicates that the relationship between the possession of computers and family wealth became weaker and also that computers became more accessible over the cycles.

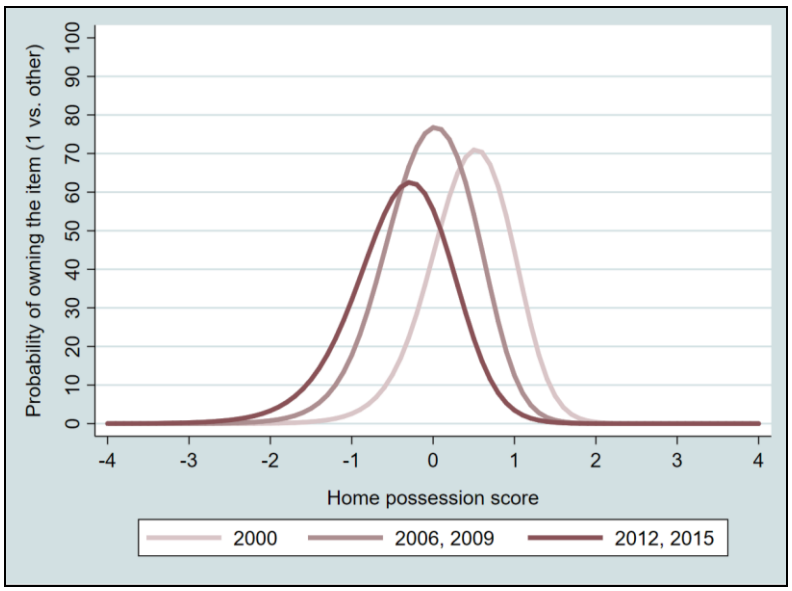


Figure 11. Model-based category response curves for computer (polytomous). For simplicity, only one category response curve is shown for each cycle.

The model-based ICCs for all the other items in the home possessions scale are presented in Appendix F.

This study is significant in that it analyzed whether and how the item parameters shifted across the cycles. The four items for which the item parameters shifted were all related to technology – educational software, internet, cell phone, and computer (polytomous). For most of these items, the relationship between the possession of the item and family wealth became weaker, or the item became more accessible over the cycles. This is not surprising, considering that technological advances made these items more affordable with time.

The results of this study can inform the selection of items to link the home possessions scale over cycles, which would make the home possessions scale longitudinally comparable, even if different subsets of items were used in different

cycles. Items that demonstrated measurement invariance over all the cycles of PISA (i.e., own room, quiet study space, school books, and dictionary) are good candidates to use as linking items.

Study 2: Measurement Invariance across Country-by-Language Groups

Figure 12 presents, for each item, the percent of country-by-language groups that required unique item parameters for the item at the end of Study 2.[24] To take into account DIF across cycles, each cycle that required unique item parameters for the item in Study 1 was counted as a separate item. It should be noted that the results are presented as percentages instead of absolute numbers because the number of country-by-language groups that administered each item depended on the number of cycles in which the item was administered as well as the number of country-by-language groups that participated in each cycle. The numerator and denominator used to calculate the percentages are presented in Appendix G.

[24] In Study 2, the item parameters were constrained to be equal across the country-by-language groups by default, and only the country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC (defined as RMSD or MD values over 0.15) were assigned unique item parameters. As a result, the model-based ICC for each country-by-language group adequately fits its observed ICC.

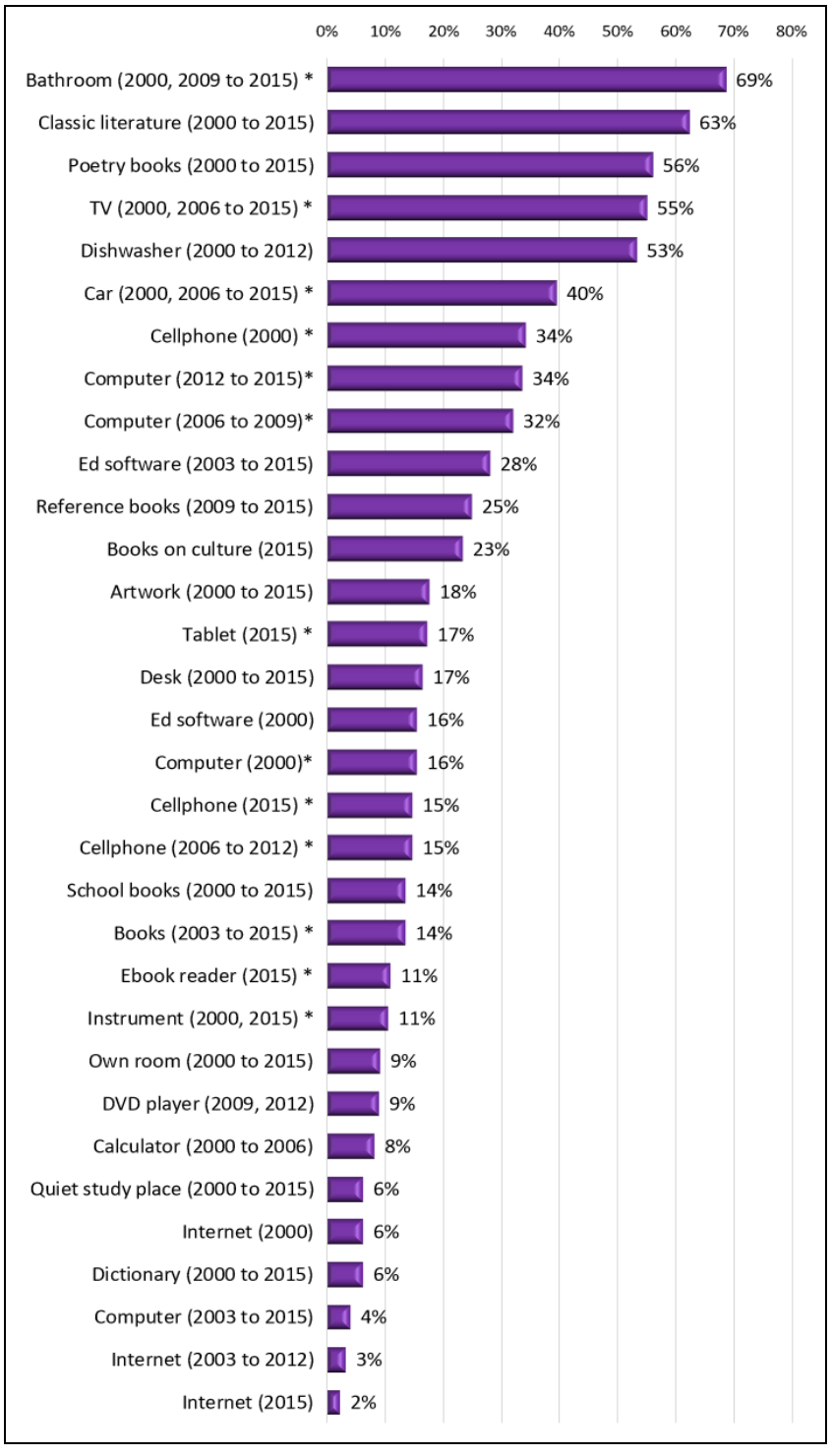


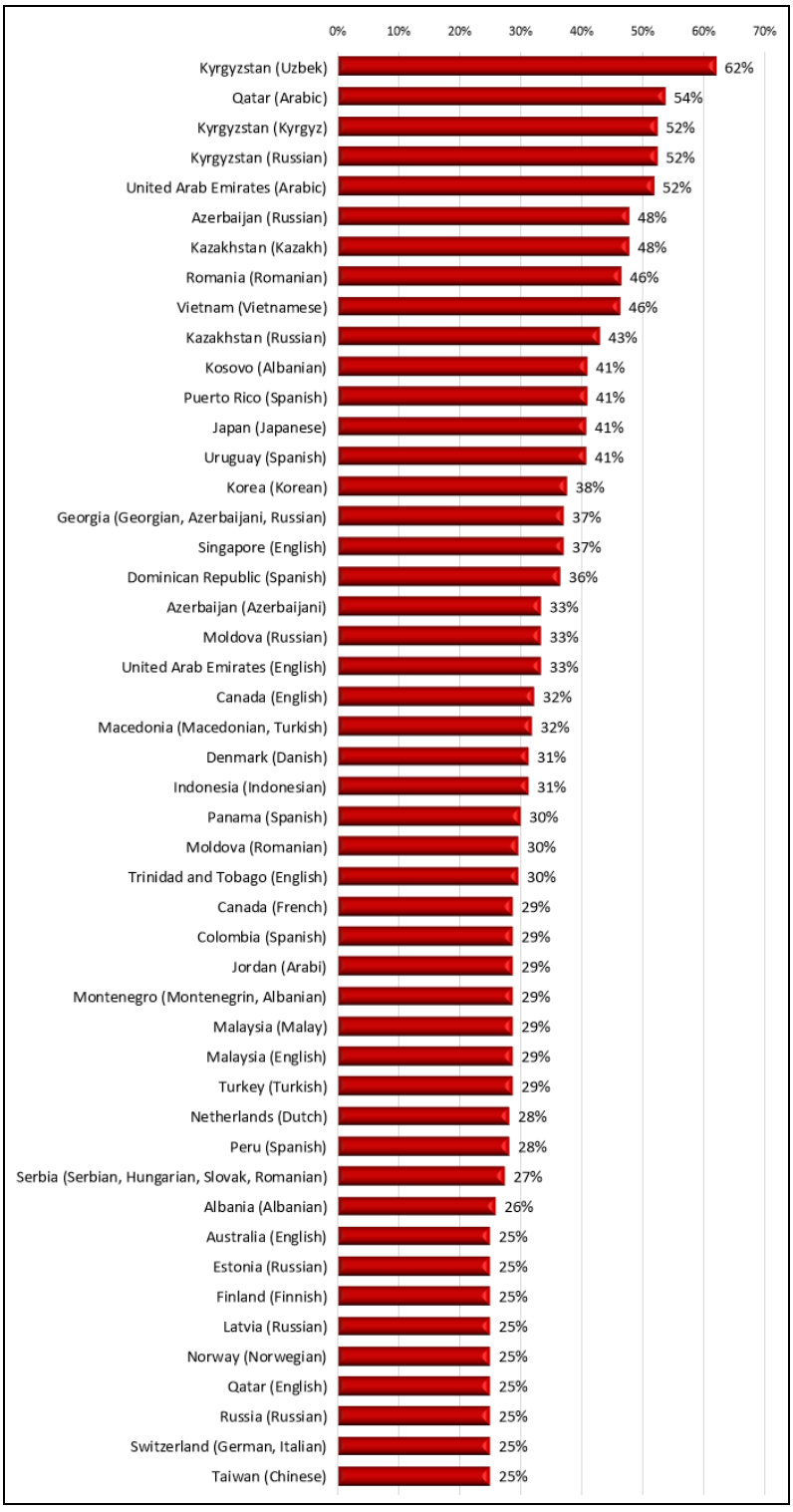
Figure 12. Percent of country-by-language groups that required unique item parameters, by item. Each cycle that required unique item parameters in Study 1 was counted as a separate item. Polytomous items are indicated with an asterisk.

Items for which over 50% of the country-by-language groups required unique item parameters included bathroom, classic literature, poetry books, TV, and dishwasher. The high percentage of country-by-language groups that required unique item parameters for these items indicated that the observed ICCs for these items were heterogenous across the country-by-language groups. In other words, these items functioned differently across the country-by-language groups when used to measure family wealth. To improve the comparability of the scale across countries in the future, it is suggested that these items be excluded from the scale.

Items for which under 10% of the country-by-language groups required unique item parameters included internet, computer (dichotomous), dictionary, quiet study place, calculator, DVD player, and own room. The low percentage of country-by-language groups that required unique item parameters for these items indicated that the observed ICCs for these items were relatively homogeneous across the country-by-language groups. In other words, these items functioned similarly across the country-by-language groups when used to measure family wealth.

The results of this study can inform the selection of items to link the home possessions scale over countries, which would make the home possessions scale comparable across the participating countries, even if different subsets of items were used in different countries. Items that demonstrated approximate measurement invariance across the country-by-language groups (i.e., own room, quiet study space, school books, and dictionary) are good candidates to use as linking items.

Results by country-by-language group. Figure 13 presents, for each country-by-language group, the percent of items that required unique item parameters at the end of Study 2. To take into account DIF across cycles, each cycle that required unique item parameters in Study 1 was counted as a separate item. Again, the results are presented as percentages instead of absolute numbers because the number of items that were administered in each country-by-language group depended on the number of cycles the country-by-language group participated in as well as the number of items that were administered in each cycle. The numerator and denominator used to calculate the percentages are presented in Appendix H.


0%
10%
20%
30%
40%
50%
60%
70%
Kyrgyzstan (Uzbek) 62%
Qatar (Arabic) 54%
Kyrgyzstan (Kyrgyz) 52%
Kyrgyzstan (Russian) 52%
United Arab Emirates (Arabic) 52%
Azerbaijan (Russian) 48%
Kazakhstan (Kazakh) 48%
Romania (Romanian) 46%
Vietnam (Vietnamese) 46%
Kazakhstan (Russian) 43%
Kosovo (Albanian) 41%
Puerto Rico (Spanish) 41%
Japan (Japanese) 41%
Uruguay (Spanish) 41%
Korea (Korean) 38%
Georgia (Georgian, Azerbaijani, Russian) 37%
Singapore (English) 37%
Dominican Republic (Spanish) 36%
Azerbaijan (Azerbaijani) 33%
Moldova (Russian) 33%
United Arab Emirates (English) 33%
Canada (English) 32%
Macedonia (Macedonian, Turkish) 32%
Denmark (Danish) 31%
Indonesia (Indonesian) 31%
Panama (Spanish) 30%
Moldova (Romanian) 30%
Trinidad and Tobago (English) 30%
Canada (French) 29%
Colombia (Spanish) 29%
Jordan (Arabi) 29%
Montenegro (Montenegrin, Albanian) 29%
Malaysia (Malay) 29%
Malaysia (English) 29%
Turkey (Turkish) 29%
Netherlands (Dutch) 28%
Peru (Spanish) 28%
Serbia (Serbian, Hungarian, Slovak, Romanian) 27%
Albania (Albanian) 26%
Australia (English) 25%
Estonia (Russian) 25%
Finland (Finnish) 25%
Latvia (Russian) 25%
Norway (Norwegian) 25%
Qatar (English) 25%
Russia (Russian) 25%
Switzerland (German, Italian) 25%
Taiwan (Chinese) 25%

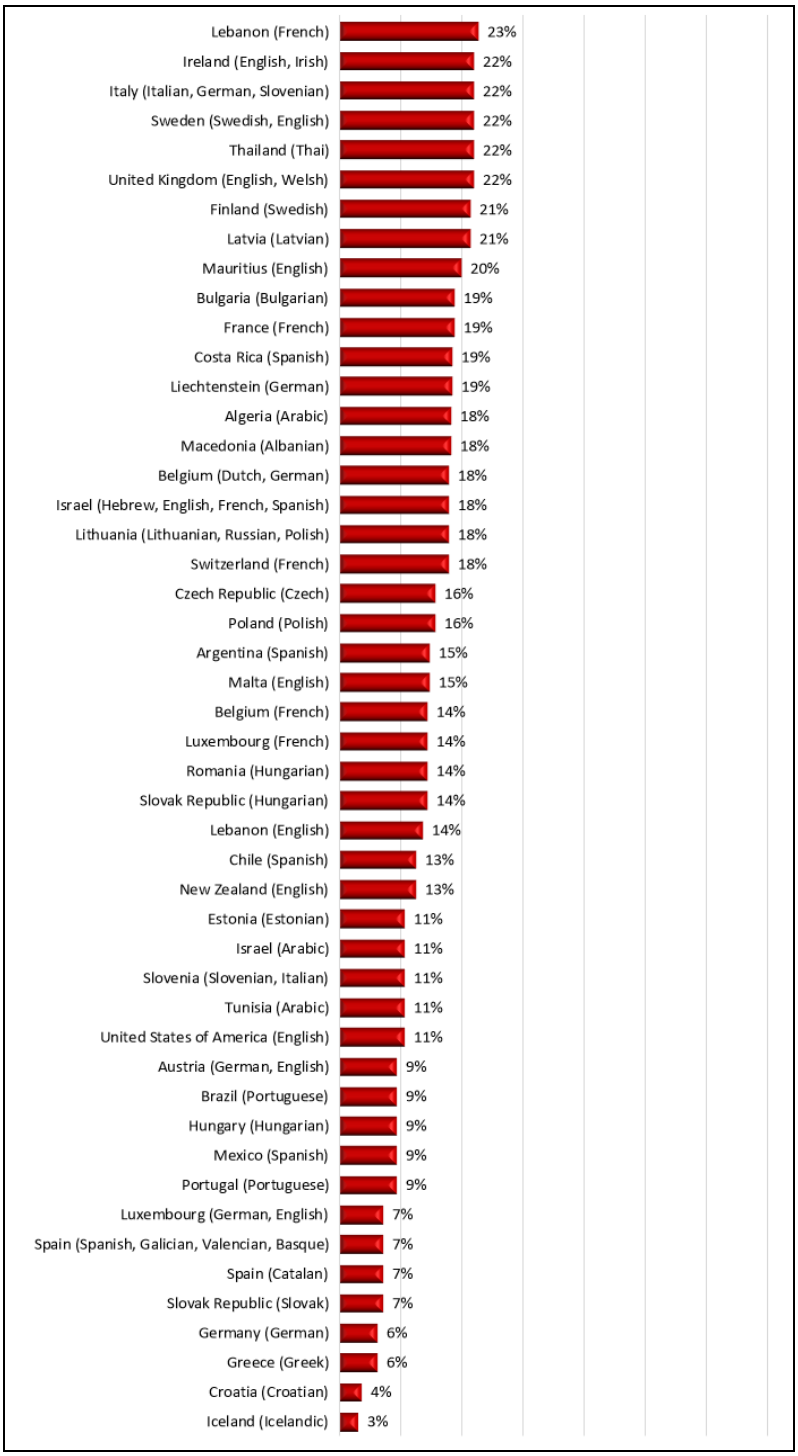


Figure 13. Percent of items that required unique item parameters, by country-by-language group. Each cycle that required unique item parameters in Study 1 was counted as a separate item.

Country-by-language groups that required unique item parameters for over 50% of the items included Kyrgyzstan (Uzbek), Qatar (Arabic), Kyrgyzstan (Kyrgyz), Kyrgyzstan (Russian), and the United Arab Emirates (Arabic). The high percentage of items that required unique item parameters in these country-by-language groups indicated that many of the items in the scale functioned differently in these country-by-language groups than in the other country-by-language groups when used to measure family wealth. In other words, in these country-by-language groups, a high level of overall misfit was found in the scale.

Country-by-language groups that required unique item parameters for under 10% of the items included Iceland (Icelandic), Croatia (Croatian), Greece (Greek), Germany (German), Spain (Spanish, Galician, Valencian, Basque), Spain (Catalan), Slovak Republic (Slovak), Luxembourg (German, English), Portugal (Portuguese), Mexico (Spanish), Hungary (Hungarian), Brazil (Portuguese), and Austria (German, English). The low percentage of items that required unique item parameters in these country-by-language groups indicated that many of the items in the scale functioned similarly in these country-by-language groups as in the other country-by-language groups when used to measure family wealth. In other words, in these country-by-language groups, a low level of overall misfit was found in the scale.

Across the country-by-language groups, the median percentage of items that required unique item parameters was 24%. In other words, half of the country-by-language groups required unique item parameters for more than 24% of the items (i.e., a relatively high level of overall misfit was found in the scale), while the other half of the

country-by-language groups required unique item parameters for less than 24% of the items (i.e., a relatively low level of overall misfit was found in the scale).

Results by region. Figure 14 presents the information in Figure 13 on a map, with country-by-language groups requiring unique item parameters for more than 24% of the items (i.e., country-by-language groups for which a relatively high level of overall misfit was found in the scale) colored in red, and country-by-language groups requiring unique item parameters for less than 24% of the items (i.e., country-by-language groups for which a relatively low level of overall misfit was found in the scale) colored in green. The purpose of this analysis was to visually inspect if there were regional differences in the overall level of misfit found in the scale.

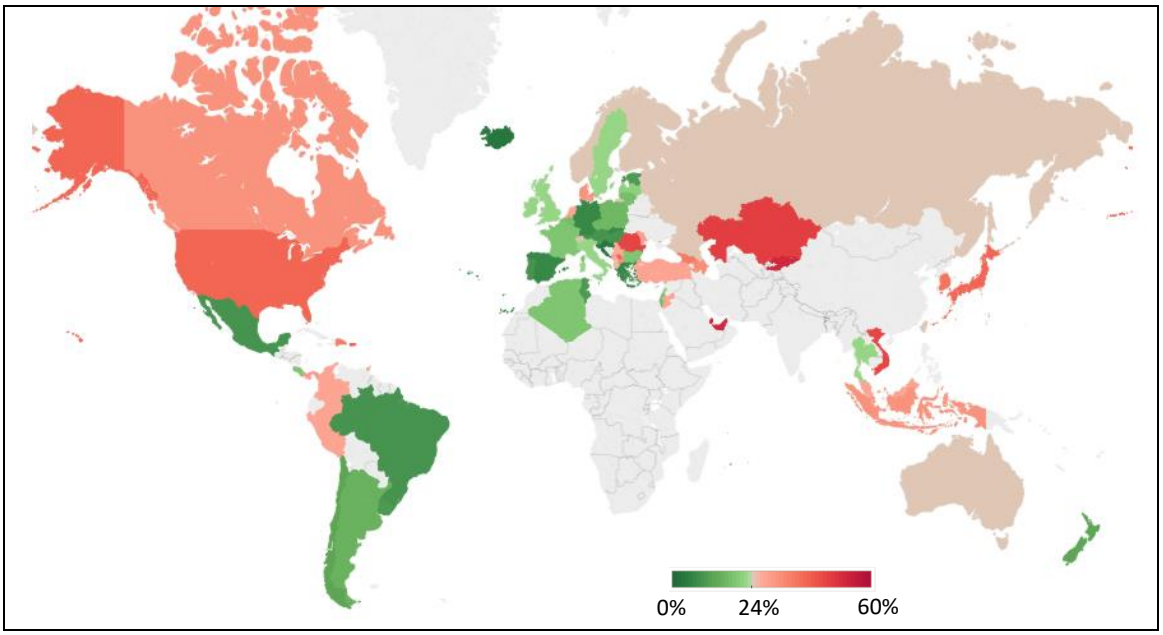


Figure 14. Percent of items that required unique item parameters, by country-by-language group. Each cycle that required unique item parameters in Study 1 was counted as a separate item. Only the main language group of each country is shown on the map.

In general, a relatively high level of overall misfit was found in the scale in North America, the Middle East (with the exception of Lebanon and Israel), Central Asia, Southeast Asia (with the exception of Thailand), and East Asia. On the contrary, a relatively low level of overall misfit was found in the scale in Western Europe (with the exception of Denmark, the Netherlands, Switzerland, Finland, and Norway) and Northern Africa. The results were mixed in Central and South America as well as in Eastern Europe. Thus, although there were some regional differences in the overall level of misfit found in the scale, exceptions were found in almost every region.

Association between the overall level of misfit and GDP per capita. The purpose of this analysis was to assess if the overall level of misfit found in the scale was associated with the country's level of economic development. Figure 15 plots each country-by-language group's average RMSD for the 22 items included in the home possessions scale in 2015 against the country's GDP per capita in 2015. The RMSDs were taken from the first round of item calibration using data only from 2015 (i.e., before any country-by-language groups received unique item parameters), and the GDP per capita is expressed in purchasing power parity (which takes into account the relative cost of living in each country).[25]

[25] Data on GDP per capita (in purchasing power parity) were obtained from the World Bank website: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?end=2015&name_desc=false&start=1960

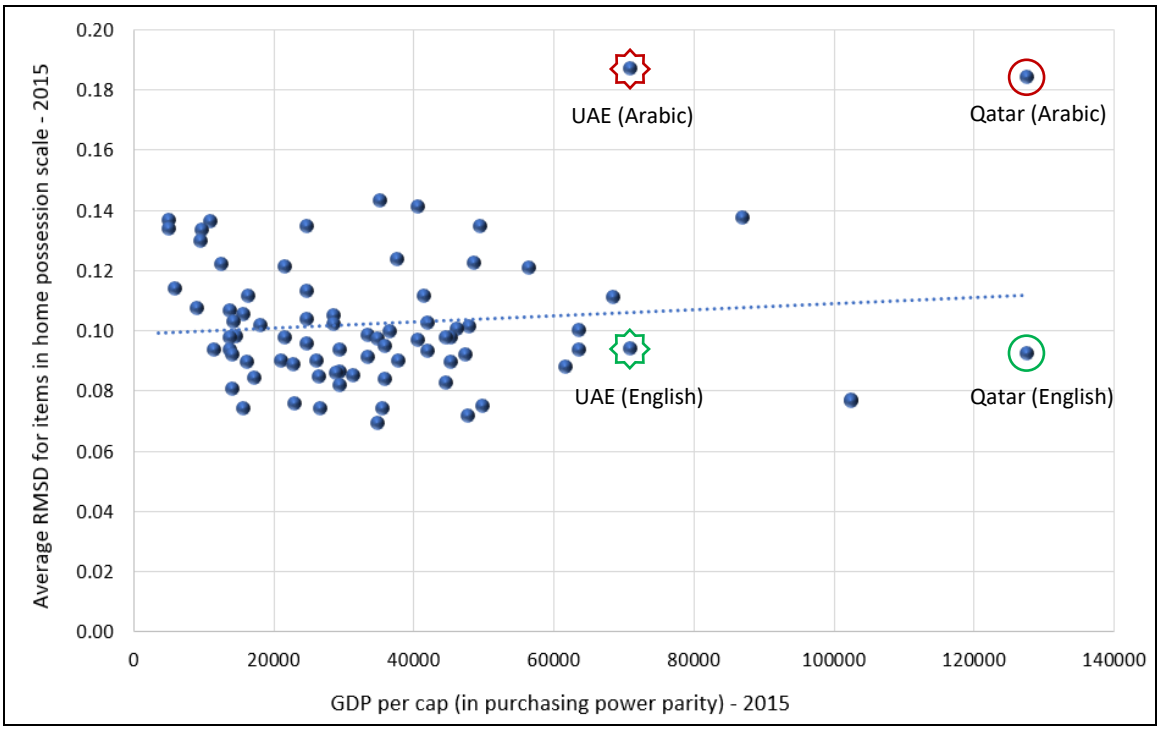


Figure 15. Scatterplot of each country-by-language group's average RMSD for the 22 items included in the home possessions scale in 2015 and the country's GDP per capita (in purchasing power parity) in 2015.

The correlation between the two variables was only 0.11, suggesting a weak relationship between the overall level of misfit in the scale and the country's level of economic development. Nevertheless, it is interesting to note that differences were detected even within countries. For example, in the United Arab Emirates and Qatar, the average RMSD was high for the Arabic-speaking group (18.7 and 18.4, respectively), while it was much lower for the English-speaking group (9.4 and 9.2, respectively). This suggests that the overall level of misfit of the home possessions scale may depend on sociocultural factors which are partially captured by the language of examination, although it may also be capturing other differences, such as the differences in response

styles (i.e., the tendency to over- or under-report home possessions) or issues with translation.

Correlation between original and new home possessions scores. Figure 16 presents a scatterplot of the original home possessions scores (obtained from the public dataset) and the new home possessions scores (generated from the final model in Study 2) for students that participated in PISA from 2003 to 2015. The 2000 cycle could not be included in the table because the public dataset did not include students' home possessions scores for this cycle. The correlation between the original and new home possessions scores was 0.90.

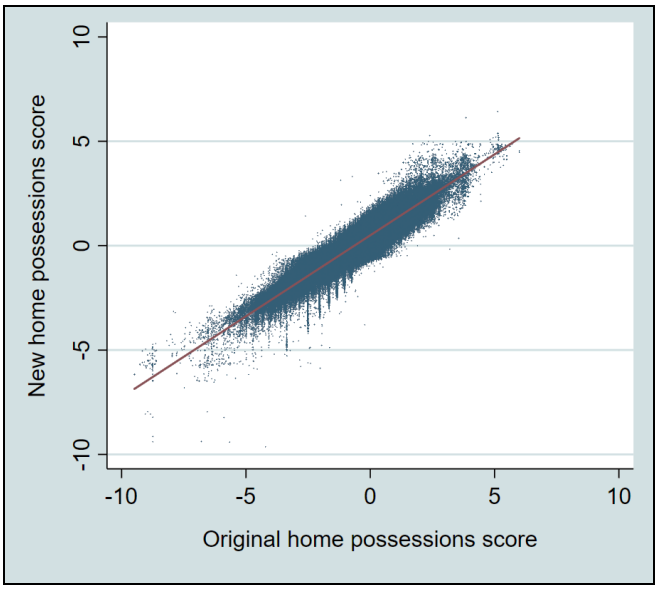


Figure 16. Scatterplot of the original home possessions scores (obtained from the public dataset) and the new home possessions scores (generated from the final model of Study 2) for students that participated in PISA from 2003 to 2015.

Table 5 presents the correlation between the original and new home possessions scores, by cycle. The correlations ranged from a minimum of 0.93 (in 2006) to a maximum of 0.97 (in 2015). It is not surprising that the correlation was the highest in 2015, since the methods used to generate the original and new home possessions scores were the most similar in 2015.

Table 5

Correlations between the Original and New Home Possessions Scores

Cycle	Correlation
2003	0.95
2006	0.93
2009	0.94
2012	0.95
2015	0.97

Study 3: Cross-Country Comparability of the New Home Possessions Scores as a Measure of SES

Figure 17 presents, for each cycle, the standard deviation of the component loadings across countries after conducting PCAs with the three components used to measure SES in PISA – home possessions, parents' education, and parents' occupation. The results for home possessions are represented by purple bars, the results for parents' education are represented by yellow bars, and the results for parents' occupation are represented by green bars. The light-colored bars indicate the results when the original

home possessions scores were included in the PCA, while the dark-colored bars indicate the results when the original home possessions scores were included in the PCA.

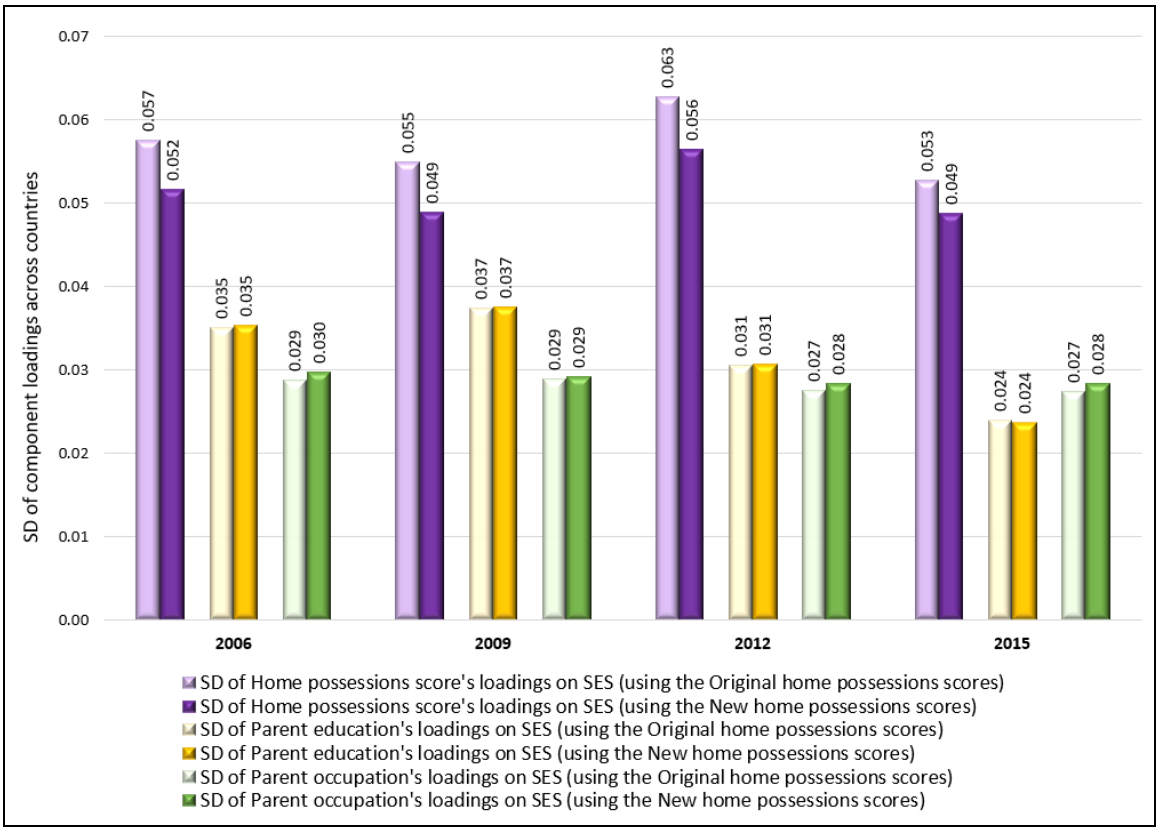


Figure 17. Standard deviation of the component loadings across countries, using the original and new home possessions scores.

As hypothesized, for the cycles from 2006 to 2012, the standard deviation of the component loadings for home possessions on SES was lower when the new home possessions scores were used than when the original home possessions scores were used. This indicated that the relationship between home possessions and SES was more consistent across countries when the new home possessions scores were used, implying that the new home possessions scores were a more comparable measure of SES across

countries. The increase in the comparability of the new home possessions scores across countries may have been due to the fact that in the new scale, the default was to constrain the item parameters to be equal across the country-by-language groups, and only the country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC were assigned unique item parameters. This is in contrast to the original method used to scale the items for these cycles, which estimated item parameters separately for each country.

In 2015, the standard deviation of the component loadings for home possessions on SES was still lower when the new home possessions scores were used than when the original home possessions scores were used. These results were unexpected because it had been hypothesized that using a lower cutoff to detect DIF in the new scale would decrease the comparability of the new home possessions scores as a measure of SES across countries.

It is also interesting to note that for both parents' education and parents' occupation, the standard deviation of the component loadings across countries changed very little when the new home possessions scores were used instead of the original home possessions scores. In addition, compared to both parents' education and parents' occupation, home possessions had the highest variability in the component loadings across countries, even when the new home possessions scores were used. This implied that among the three components used to measure SES in PISA, home possessions had the most heterogenous relationship with SES across countries. Also, among the three components used to measure SES, home possessions had the weakest relationship with

SES, indicated by the lowest average component loading on SES, as presented in Table 6. This suggests that more improvements should be made to the home possessions scale so that it has a stronger relationship as well as a more stable relationship with SES across countries.

Table 6

Average Component Loadings across Countries, Using the Original and New Home Possessions Scores

	Average component loadings across countries (SD)		
	Home possessions	Parents' education	Parents' occupation
PCA using original home possessions scores in 2006	0.77 (0.057)	0.83 (0.035)	0.81 (0.029)
PCA using new home possessions scores in 2006	0.78 (0.052)	0.83 (0.035)	0.81 (0.030)
PCA using original home possessions scores in 2009	0.77 (0.055)	0.83 (0.037)	0.82 (0.029)
PCA using new home possessions scores in 2009	0.79 (0.049)	0.84 (0.037)	0.81 (0.029)
PCA using original home possessions scores in 2012	0.77 (0.063)	0.83 (0.031)	0.82 (0.027)
PCA using new home possessions scores in 2012	0.77 (0.056)	0.83 (0.031)	0.82 (0.028)
PCA using original home possessions scores in 2015	0.78 (0.053)	0.82 (0.024)	0.82 (0.027)
PCA using new home possessions scores in 2015	0.77 (0.049)	0.82 (0.024)	0.81 (0.028)

Study 4: Predicting PISA Cognitive Scores with the Original and New Home Possessions Scores

Figure 18 presents the average r^2 across countries of the bivariate regressions predicting students' scores on the PISA cognitive assessments with the home possessions scores. The r^2 for math is represented by red bars, the r^2 for science is represented by blue bars, and the r^2 for reading is represented by green bars. For each cycle, the r^2 using the original home possessions scores is indicated by light-colored bars, while the r^2 using the new home possessions scores is indicated by dark-colored bars. It should be noted that the United States of America was excluded from the analysis predicting students' reading scores in 2006 because the public dataset did not include students' reading scores for this country in 2006.

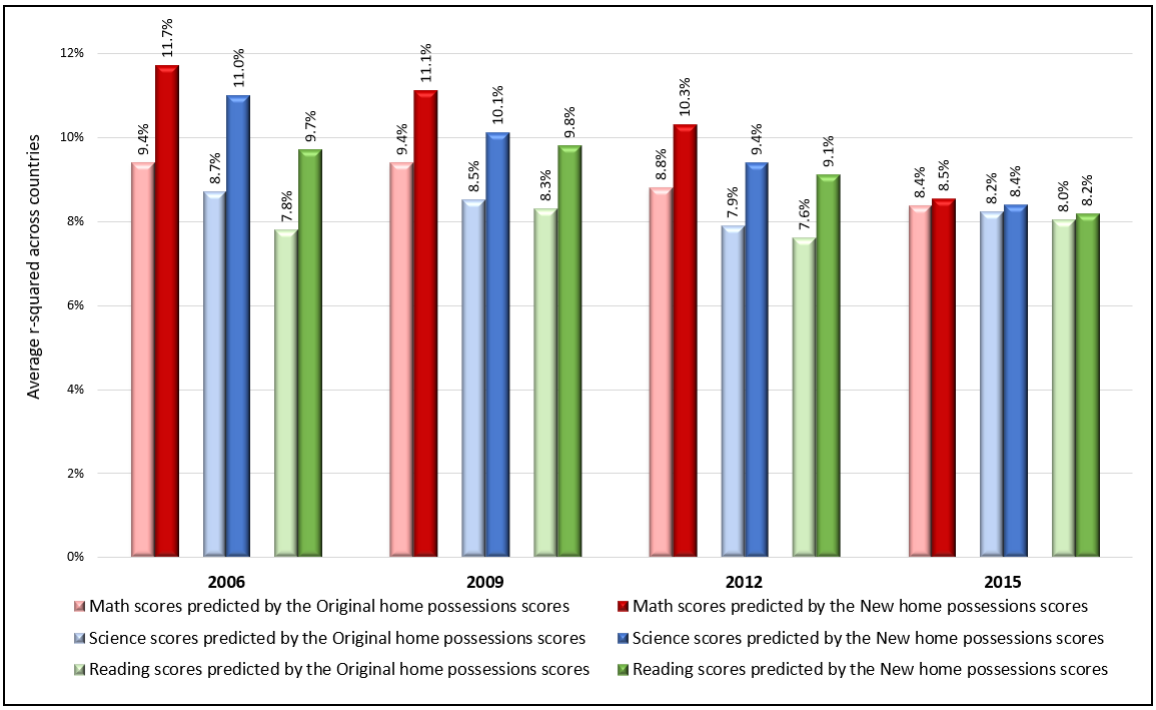


Figure 18. Average r^2 across countries of the bivariate regressions predicting students' scores on the PISA cognitive assessments with the home possessions scores.

As hypothesized, for the cycles from 2006 to 2012, the new home possessions scores explained more of the variation in the PISA cognitive scores than the original home possessions scores for all three subjects. In other words, the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores, which implied that the new home possessions scores were a more accurate measure of SES than the original home possessions scores within countries. While the accuracy of the new home possessions scores may have decreased for these cycles because the item parameters were constrained to be equal across the country-by-language groups by default (instead of having the item parameters estimated separately for each country), the accuracy of the new home possessions scores may have been

greatly increased by using the 2PL model and the GPCM to calibrate the items (instead of the Rasch model and PCM, respectively).

As hypothesized, for the 2015 cycle, the new home possessions scores explained as much of the variation in the PISA cognitive scores as the original home possessions scores for all three subjects.[26] In other words, the predictive power of the new home possessions scores and the original home possessions scores were similar, which implied that the new home possessions scores were as accurate as the original home possessions scores in measuring SES. While the accuracy of the new home possessions scores for this cycle may have increased because more country-by-language groups were assigned unique item parameters, this may have been balanced out by calibrating the items parameters with data from all cycles instead of using data only from the 2015 cycle.

The results of this study provide important evidence that the new home possessions scores are at least as accurate as the original home possessions scores in measuring SES within countries, even though the new home possessions scores are also a more comparable measure of SES across countries. In other words, the accuracy of scores within countries was not compromised by the increase in the comparability of the scores across countries.

For reference, Table 7 presents the bivariate correlations between the three components used to measure SES in PISA – home possessions scores (from the new model), parents' education, and parents' occupation. In each cycle, the correlations are the weakest between the home possessions scores and parents' education (ranging from

[26] While there were slight differences in the r^2, the differences were not statistically significant.

0.44 to 0.49), while the correlations are the strongest between parents' education and parents' occupation (ranging from 0.51 to 0.55).

Table 7

Correlations between the New Home Possessions Scores, Parents' Education, and Parents' Occupation, by cycle

	Cycle			
	2006	2009	2012	2015
New home possessions scores & Parents' education	0.47	0.49	0.47	0.44
New home possessions scores & Parents' occupation	0.44	0.44	0.44	0.43
Parents' education & Parents' occupation	0.52	0.53	0.55	0.51

Figure 19 presents the standardized regression coefficients for multivariable regressions predicting students' cognitive scores on PISA with the new home possessions scores, parents' education, and parents' occupation. The results for the regressions predicting math are represented by red bars, the results for the regressions predicting science are represented by blue bars, and the results for the regressions predicting science are represented by green bars. In every cycle, the lowest standardized regression coefficient is for parents' education. As for the highest standardized regression coefficient, in 2006, it is home possessions; in 2009, it is either home possessions or parents' occupation, depending on the subject; and in 2012 and 2015, it is parents' occupation.

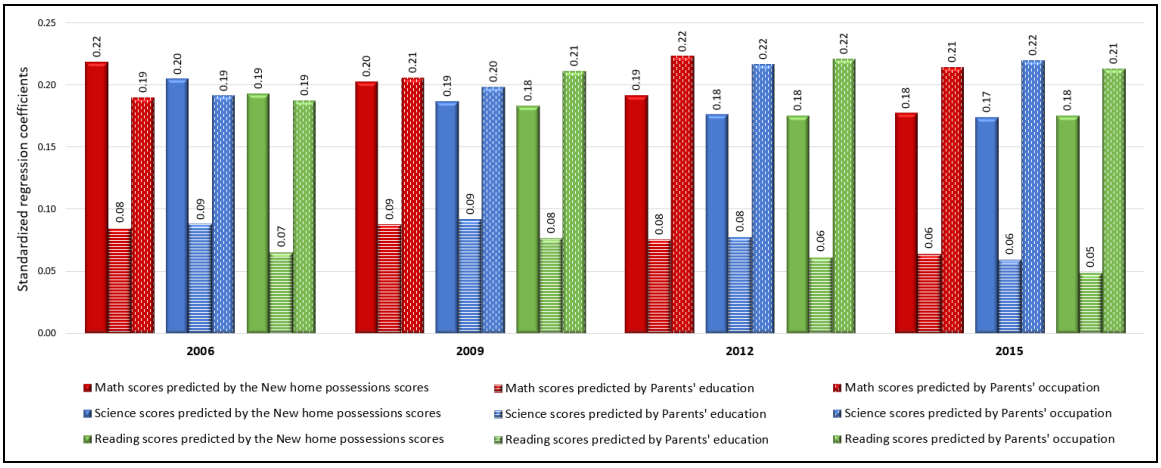


Figure 19. Standardized regression coefficients for models predicting students' cognitive scores on PISA with the new home possessions scores, parents' education, and parents' occupation.

Study 5: Evidence Supporting the External Validity of the New Home Possessions Scores as a Measure of SES

Figure 20 presents the correlation between countries' average new home possessions score and HDI, by cycle. The correlations using the original home possessions scores are indicated by light-colored bars, while the correlations using the new home possessions scores are indicated by dark-colored bars.

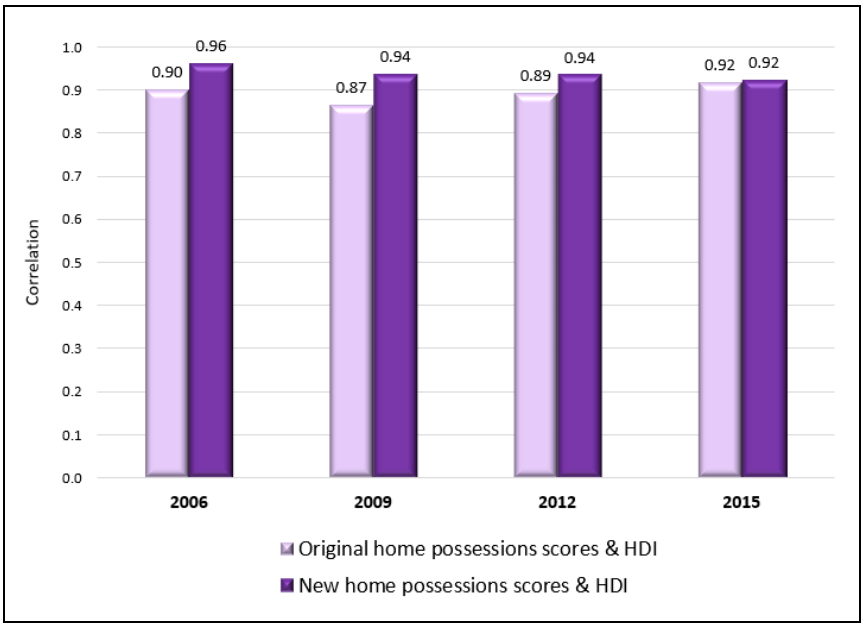


Figure 20. Correlation between countries' average home possessions score and HDI.

For every cycle, the correlation between countries' average new home possessions score and HDI was over 0.90, representing a strong correlation between the two variables. Since HDI and SES are measured using similar components, the strong correlation between the average new home possessions scores and HDI was taken as evidence supporting the external validity of the new home possessions scores as a measure of SES.

It is interesting to note that for the cycles from 2006 to 2012, the correlation between countries' average home possessions score and HDI was higher when the new home possessions scores were used instead of the original home possessions scores. This was most likely because the new home possessions scores were a more accurate measure of SES than the original home possessions scores, for reasons explained above. In 2015, the correlation between countries' average home possessions score and HDI was similar

when either the original or new home possessions scores were used. Again, for reasons explained above, this was most likely because the new home possessions scores were as accurate as the original home possessions scores in measuring SES.

CHAPTER 4 – CONCLUSION

Significance of the Study

Measuring SES is very important in educational research because this information is often used by researchers to contextualize the results of an assessment or to control for SES when analyzing the relationship between academic achievement and other variables. While most of the research on family SES and students' academic achievement have been conducted in developed countries, international large-scale assessments, such as PISA, TIMSS, and PIRLS, have made it possible to conduct such research in a wide range of countries as well as to make cross-country comparisons. The interest in analyzing the relationship between SES and educational achievement has also increased among educational policy makers, due to the increased focus on equity in the global education agenda, such as the Sustainable Development Goals and the Education 2030 Framework for Action. However, as noted by Rutkowski and Rutkowski (2013), any cross-country comparisons using SES data from international large-scale assessments should be preceded by a careful examination of the psychometric properties of the scale used to measure SES, a topic which is rarely addressed by researchers.

The current study was designed to fill the gaps in this area of research by analyzing the longitudinal and cross-country measurement invariance of the PISA home possessions scale which is one of the three components, along with parents' education and parents' occupation, used to measure students' SES in PISA. This scale was developed for the first cycle of PISA in 2000 when only 43 countries participated, of

which two-thirds were members of the OECD. Very little changes have been made to the scale since then, even though in the most recent cycle of PISA in 2018, 79 countries participated, of which less than half were members of the OECD. With more countries planning to participate in future cycles of PISA, the participating countries will become increasingly heterogeneous, presenting further challenges to the comparability of the home possessions scores across countries.

In this research, Study 1 analyzed the longitudinal measurement invariance of the items included in the PISA home possessions scale. Although trend analyses for this scale have been conducted in the past (OECD, 2012, p. 314; OECD, 2014b, p. 353; OECD, 2017, p. 342), there is a lack of documentation in the PISA technical reports on the methodology and results of these analyses, making it impossible for researchers to assess the methods or to conduct further analyses with the results. The current study found that among the 25 items included in the home possessions scale, longitudinal measurement invariance could not be established for four items, all related to technology – educational software, internet, cell phone, and computer (polytomous). This study is significant because it revealed that many of the items in the scale were invariant over time. Also, the results of this study can inform the selection of items to link the home possessions scale over the PISA cycles, which would make the home possessions scale longitudinally comparable even if different subsets of items were used in different cycles.

Study 2 analyzed the measurement invariance of the items across country-by-language groups, another area that lacked documentation in the PISA technical reports. The study found that some items in the scale (i.e., bathroom, classic literature, poetry

books, TV, and dishwasher) functioned differently across the country-by-language groups when used to measure family wealth. To improve the comparability of the scale across countries in the future, it is suggested that these items be excluded from the scale. The study also found some items (i.e., internet, computer-dichotomous, dictionary, and quiet place to study) were a relatively comparable measure of family wealth across the country-by-language groups. These items may be used to link the home possessions scale across different countries in the future, which would make the home possessions scale comparable across the participating countries even if different subsets of items were used in different countries.

Another finding of Study 2 was that in some country-by-language groups – such as Kyrgyzstan (Uzbek), Qatar (Arabic), Kyrgyzstan (Kyrgyz), Kyrgyzstan (Russian), and the United Arab Emirates (Arabic) – many of the items functioned differently than in other country-by-language groups when used to measure family wealth. The overall level of misfit found in the scale was not associated with the country's GDP per capita, while some evidence suggested that it may be associated with sociocultural factors (which were partially captured by the language of examination) and the region in which the country was located. Further research should investigate other country-level factors that are associated with the overall level of misfit found in the scale, which may help to improve the cross-country comparability of the scale in the future.

In Study 3, PCAs were conducted with home possessions, parents' education, and parents' occupation – the three components used to measure SES in PISA. The study found that for all of the cycles included in Study 3, the component loadings of home

possessions on SES had lower variability across countries when the new home possessions scores (generated from the final model in Study 2) were included in the PCA instead of the original home possessions scores (obtained from the public dataset). This implied that the new home possessions scores were a more comparable measure of SES across countries than the original home possessions scores. The increase in the cross-country comparability of the new home possessions scores may have been a result of constraining the item parameters to be equal across the country-by-language groups by default (and assigning unique item parameters only to the country-by-language groups for which the observed ICC exhibited substantial misfit with the international ICC).

In Study 4, it was found that for most of the cycles included in the study, the new home possessions scores were a better predictor of the PISA cognitive scores than the original home possessions scores, while in Study 5, it was found that the correlation between countries' average home possessions score and HDI was higher when the new home possessions scores were used instead of the original home possessions scores. These results implied that the new home possessions scores were a more accurate measure of SES within countries than the original home possessions scores, even though some of the methods used to increase the comparability of the home possessions scores across countries decreased the accuracy of the scores within countries. The increase in the accuracy of the new home possessions scores may have been a result of using the 2PL model and the GPCM to calibrate the items (instead of the Rasch model and PCM, respectively), which can model the relationship between the possession of an item and family wealth more accurately.

Study 5 also found that for all cycles included in the analysis, the correlation between countries' average new home possessions score and HDI was over 0.90, representing a strong correlation between the two variables. Since HDI and SES are measured using similar components, the strong correlations between the average new home possessions scores and HDI provided evidence supporting the external validity of the new home possessions scores as a measure of SES.

Limitations of the Study

As with any research, the findings of this study should be interpreted in light of its limitations. The first limitation is that this study relied on self-reported data from the PISA student questionnaires. While the study assumed that all students had replied conscientiously and accurately to the questionnaire, this may be an unrealistic assumption. In fact, Akyol, Krishna, and Wang (2018) used data on the time that students spent on each item of PISA 2015 as well as the way in which they responded to various types of items, and came to the conclusion that many students did not take the PISA assessment seriously. The percentage of non-serious test takers ranged from 14% in Korea (South) to 67% in Brazil. Although the study only analyzed data from the cognitive assessments, it sheds light into how seriously students respond to low-stakes assessments such as PISA. Even for students who take PISA seriously, the accuracy of their responses may have deteriorated in the later sections of the survey due to respondent fatigue (Ben-Nun, 2008). Considering that students had to take the PISA cognitive assessment for two hours before responding to the student questionnaire (in which they

had to respond to up to 220 items in 35 minutes), respondent fatigue may have affected the accuracy of their responses. Although it is impossible to check how conscientiously and accurately the students responded to the PISA home possessions scale, an analysis of the two items regarding computers (i.e., a polytomous item asking how many computers the student had at home, and a dichotomous item asking if the student had a computer at home that he or she could use for school work) provides some insights. Pooling data across all cycles and restricting the sample to those who responded to both items, among the students who responded that they did not have a computer at home, 7.9% responded that they had a computer at home which they could use for school work, raising some doubts about the reliability of students' self-reports on the home possessions scale.

Second, all countries with a sizeable minority language population were excluded from the dataset in 2000 and 2003 because there was no information in the public dataset on the language of examination for these cycles, making it impossible to divide these countries into country-by-language groups for Study 2. As a result, 10 countries (out of 43 countries) were excluded from the final dataset in 2000, while nine countries (out of 41 countries) were excluded from the final dataset in 2003. This may have affected the observed ICCs for 2000 and 2003 as well as the total sample ICC (which was estimated with data from all cycles). However, for most of the items, the item parameters estimated with data from all countries that participated in 2000 and 2003 were very similar to the item parameters estimated with data excluding countries with a sizeable minority language population, as presented in Appendix I and Appendix J. Therefore, it can be deduced that excluding the countries with a sizeable minority language population from

the dataset in 2000 and 2003 did not have a large effect on the observed ICCs for 2000 and 2003 as well as the total sample ICC.

Third, the language of examination was used to divide students within a country into subgroups, based on the assumption that the relationship between the possession of an item and family wealth may be different for different groups, due to sociocultural reasons (Brese & Mirazchiyski, 2013; Yang & Gustafsson, 2004). However, if the language in which a student is assessed is not reflective of the sociocultural group to which the student belongs, using the language of examination will not be an effective way to divide the population into different sociocultural groups. As a case in point, almost a quarter of Peruvians identify themselves as Quechua (National Institute of Statistics and Informatics of Peru, 2018), an indigenous group with a distinct culture and language. However, the PISA assessment is only offered in Spanish in Peru, masking the different sociocultural groups within the country. The same can be said of other multicultural countries that offer the PISA assessment in only one language, such as the United States of America and Singapore. Nevertheless, in the absence of other indicators that can be used to divide students into different sociocultural groups (i.e., language spoken at home, ethnicity, and religion), the language of examination may be the best alternative.

Fourth, due to the study design, it was not possible to analyze the extent to which different methods affected the comparability of the new home possessions scores across countries and the accuracy of the new home possessions scores within countries. While the new home possessions scores were found to be a more comparable measure of SES

across countries than the original home possessions scores for all cycles included in Study 3, it was not clear how much of the increase in comparability was due to constraining the item parameters to be equal across the country-by-language groups by default, and how much of the decrease in comparability was due to using a lower cutoff to detect DIF. Similarly, while the new home possessions scores were found to be a more accurate measure of SES within countries than the original home possessions scores for most of the cycles included in Study 4 and 5, it was not clear how much of the increase in accuracy was due to using the 2PL model and the GPCM to calibrate the items (instead of the Rasch model and PCM, respectively) and by using a lower cutoff to detect DIF, and how much of the decrease in accuracy was due to constraining the item parameters to be equal across the country-by-language groups by default and by calibrating the item parameters with data from all cycles (instead of calibrating the item parameters separately for each cycle). To analyze how a particular method affects the comparability and accuracy of the home possessions scores, future research should include simulation studies in which only one change is made to the method at a time.

Lastly, assigning unique item parameters to certain country-by-language groups may have resulted in parameters that conformed to the random variability of the sample, especially for small samples. However, considering that all of the country-by-language groups had a sample size of at least 1,000 students, with the exception of the Azerbaijan-Russian speaking group (which had a sample size of 473 students), it was assumed that the country-by-language groups would be large enough to make the item parameter

estimates robust to the idiosyncrasies of the samples. Appendix K presents a histogram of the sample size of the country-by-language groups.

In spite of these limitations, this study provided many insights into the longitudinal and cross-country measurement invariance of the PISA home possessions scale. Also, the study generated home possessions scores that were a more comparable measure of SES across countries than the original home possessions scores, while the accuracy of the scores in measuring SES within countries was maintained or improved. The results of this study can also inform the selection of items that can be used to link the home possessions scale over cycles and across countries, so the comparability of the scale can be maintained even when different subsets of items are used in different cycles or countries. In sum, this study can help improve the PISA home possessions scale, so it can continue to provide valuable information to researchers and policy makers on SES over the PISA cycles and across the countries that participate in PISA.

APPENDICES

Appendix A:

Countries Included in the Final Dataset

	Total sample size (unweighted)							# of cy -cles
	2000	2003	2006	2009	2012	2015	Total	
Albania	2,783			4,596	4,743		12,122	3
Algeria						5,519	5,519	1
Argentina	2,230		4,339	4,774	5,908		17,251	4
Australia	2,859	12,551	14,170	14,251	14,481	14,530	72,842	6
Austria	2,640	4,597	4,927	6,590	4,755	7,007	30,516	6
Azerbaijan			5,184	4,691			9,875	2
Belgium			8,857	8,501	8,597	9,651	35,606	4
Brazil	2,717	4,452	9,295	20,127	19,204	23,141	78,936	6
Bulgaria	2,615		4,498	4,507	5,282	5,928	22,830	5
Canada			22,646	23,207	21,544	20,058	87,455	4
Chile	2,721		5,233	5,669	6,856	7,053	27,532	5
Colombia			4,478	7,921	9,073	11,795	33,267	4
Costa Rica				4,578	4,602	6,866	16,046	3
Croatia			5,213	4,994	5,008	5,809	21,024	4
Czech Republic	3,066	6,320	5,932	6,064	5,327	6,894	33,603	6
Denmark	2,382	4,218	4,532	5,924	7,481	7,161	31,698	6
Dominican Republic						4,740	4,740	1
Estonia			4,865	4,727	4,779	5,587	19,958	4
Finland			4,714	5,810	8,829	5,882	25,235	4
France	2,597	4,300	4,716	4,298	4,613	6,108	26,632	6

Georgia				4,646		5,316	9,962	2
Germany	2,830	4,660	4,891	4,979	5,001	6,504	28,865	6
Greece	2,605	4,627	4,873	4,969	5,125	5,532	27,731	6
Hungary	2,799	4,765	4,490	4,605	4,810	5,658	27,127	6
Iceland	1,882	3,350	3,789	3,646	3,508	3,371	19,546	6
Indonesia	4,089	10,761	10,647	5,136	5,622	6,513	42,768	6
Ireland	2,128	3,880	4,585	3,937	5,016	5,741	25,287	6
Israel			4,584	5,761	5,055	6,598	21,998	4
Italy	2,765	11,639	21,773	30,905	31,073	11,583	109,738	6
Japan	2,924	4,707	5,952	6,088	6,351	6,647	32,669	6
Jordan			6,509	6,486	7,038	7,267	27,300	4
Kazakhstan				5,412	5,808		11,220	2
Korea (South)	2,769	5,444	5,176	4,989	5,033	5,581	28,992	6
Kosovo						4,826	4,826	1
Kyrgyzstan			5,904	4,986			10,890	2
Latvia			4,719	4,502	4,306	4,869	18,396	4
Lebanon						4,546	4,546	1
Liechtenstein	175	332	339	329	293		1,468	5
Lithuania			4,744	4,528	4,618	6,525	20,415	4
Luxembourg			4,567	4,622	5,258	5,299	19,746	4
Macedonia						5,324	5,324	1
Malaysia				4,999	5,197		10,196	2
Malta				3,453		3,634	7,087	2
Mauritius				4,654			4,654	1
Mexico	2,567	29,983	30,971	38,250	33,806	7,568	143,145	6
Moldova				5,194		5,325	10,519	2

Montenegro			4,455	4,825	4,744	5,665	19,689	4
Netherlands	1,382	3,992	4,871	4,760	4,460	5,385	24,850	6
New Zealand	2,048	4,511	4,823	4,643	4,291	4,520	24,836	6
Norway	2,307	4,064	4,692	4,660	4,686	5,456	25,865	6
Panama				3,969			3,969	1
Peru	2,460			5,985	6,035	6,971	21,451	4
Poland	1,976	4,383	5,547	4,917	4,607	4,478	25,908	6
Portugal	2,545	4,608	5,109	6,298	5,722	7,325	31,607	6
Puerto Rico						1,398	1,398	1
Qatar			6,265	9,078	10,966	12,083	38,392	4
Romania			5,118	4,776	5,074	4,876	19,844	4
Russia	3,719	5,974	5,799	5,308	5,231	6,036	32,067	6
Serbia			4,798	5,523	4,684		15,005	3
Singapore				5,283	5,546	6,115	16,944	3
Slovak Republic			4,731	4,555	4,678	6,350	20,314	4
Slovenia			6,595	6,155	5,911	6,406	25,067	4
Spain			19,604	25,887	25,313	6,736	77,540	4
Sweden	2,464	4,624	4,443	4,567	4,736	5,458	26,292	6
Switzerland			12,192	11,812	11,229	5,860	41,093	4
Taiwan			8,815	5,831	6,046	7,708	28,400	4
Thailand	2,959	5,236	6,192	6,225	6,606	8,249	35,467	6
Trinidad and Tobago			4,778			4,692	9,470	2
Tunisia		4,721	4,640	4,955	4,407	5,375	24,098	5
Turkey		4,855	4,942	4,996	4,848	5,895	25,536	5
United Arab Emirates			10,867		11,500	14,167	36,534	3
United Kingdom	5,195	9,535	13,152	12,179	12,659	14,157	66,877	6

USA	2,135	5,456	5,611	5,233	4,978	5,712	29,125	6
Uruguay		5,835	4,839	5,957	5,315	6,062	28,008	5
Vietnam					4,959	5,826	10,785	2
Total	83,333	188,380	389,345	492,327	463,231	456,917	2,073,533	
Total # of countries	32	30	55	68	61	65	75	

Appendix B:

Percent of the Sample with Data on Each Item of the Home Possessions Scale

	Cycle					
	2000	2003	2006	2009	2012	2015
Desk	98 %	99 %	98 %	98 %	97 %	97 %
Own room	98 %	99 %	98 %	98 %	96 %	96 %
Quiet study place	98 %	99 %	98 %	98 %	97 %	97 %
Computer		99 %	98 %	98 %	98 %	97 %
Ed software	97 %	99 %	96 %	**97 %**	**95 %**	95 %
Internet	97 %	99 %	98 %	98 %	97 %	97 %
Classic literature	97 %	99 %	97 %	97 %	96 %	96 %
Poetry books	97 %	99 %	97 %	98 %	96 %	96 %
Artwork	98 %	99 %	97 %	97 %	96 %	96 %
School books	98 %	99 %	98 %	98 %	97 %	97 %
Reference books				97 %	95 %	95 %
Dictionary	98 %	99 %	98 %	98 %	98 %	97 %
Books on culture						96 %
Calculator	98 %	99 %	98 %			
Dishwasher	98 %	**95 %**	**95 %**	98 %	97 %	
DVD player				98 %	98 %	
TV *	98 %		99 %	99 %	98 %	97 %
Car *	97 %		98 %	98 %	97 %	96 %
Bathroom *	98 %			98 %	96 %	**95 %**
Cellphone *	97 %		99 %	99 %	98 %	97 %
Computer *	**97 %**		98 %	98 %	98 %	97 %
Tablet *						97 %
Ebook reader *						96 %
Instrument *	97 %					97 %
Books *		97 %	98 %	98 %	97 %	98 %

Note. Polytomous items are indicated with an asterisk. The smallest value for each cycle is indicated in bold.

Appendix C:

Language Groups in Each Country

Country	Language group 1 (% of sample using weights, if not 100%)[a]	Language group 2 (% of sample using weights)	Language group 3 (% of sample using weights)
Albania	Albanian		
Algeria	Arabic		
Argentina	Spanish		
Australia	English		
Austria	German, English		
Azerbaijan	Azerbaijani (95%)	Russian (5%)	
Belgium	Dutch, German (57%)	French (43%)	
Brazil	Portuguese		
Bulgaria	Bulgarian		
Canada	English (77%)	French (23%)	
Chile	Spanish		
Colombia	Spanish		
Costa Rica	Spanish		
Croatia	Croatian		
Czech Republic	Czech		
Denmark	Danish		
Dominican Republic	Spanish		
Estonia	Estonian (79%)	Russian (21%)	
Finland	Finnish (94%)	Swedish (6%)	
France	French		
Georgia	Georgian, Azerbaijani, Russian		
Germany	German		
Greece	Greek		
Hungary	Hungarian		
Iceland	Icelandic		
Indonesia	Indonesian		
Ireland	English, Irish		
Israel	Hebrew, English, French, Spanish (79%)	Arabic (21%)	
Italy	Italian, German, Slovenian		
Japan	Japanese		
Jordan	Arabic		
Kazakhstan	Kazakh (62%)	Russian (38%)	

Korea (South)	Korean		
Kosovo	Albanian		
Kyrgyzstan	Kyrgyz (65%)	Russian (26%)	Uzbek (9%)
Latvia	Latvian (76%)	Russian (24%)	
Lebanon	French (63%)	English (37%)	
Liechtenstein	German		
Lithuania	Lithuanian, Russian, Polish		
Luxembourg	German, English (75%)	French (25%)	
Macedonia	Macedonian, Turkish (75%)	Albanian (25%)	
Malaysia	Malay (87%)	English (13%)	
Malta	English		
Mauritius	English		
Mexico	Spanish		
Moldova	Romanian (81%)	Russian (19%)	
Montenegro	Montenegrin, Albanian		
Netherlands	Dutch		
New Zealand	English		
Norway	Norwegian		
Panama	Spanish		
Peru	Spanish		
Poland	Polish		
Portugal	Portuguese		
Puerto Rico	Spanish		
Qatar	Arabic (63%)	English (37%)	
Romania	Romanian (95%)	Hungarian (5%)	
Russia	Russian		
Serbia	Serbian, Hungarian, Slovak, Romanian		
Singapore	English		
Slovak Republic	Slovak (94%)	Hungarian (6%)	
Slovenia	Slovenian, Italian		
Spain	Spanish, Galician, Valencian, Basque (83%)	Catalan (17%)	
Sweden	Swedish, English		
Switzerland	German, Italian (75%)	French (25%)	
Taiwan	Mandarin		
Thailand	Thai		
Trinidad and Tobago	English		
Tunisia	Arabic		
Turkey	Turkish		

United Arab Emirates	Arabic (59%)	English (41%)
United Kingdom	English, Welsh	
United States of America	English	
Uruguay	Spanish	
Vietnam	Vietnamese	

Appendix D:

Percent of the Sample with Data on Home Possessions Scores, Parents' Education, and Parents' Occupation

	2006	2009	2012	2015
Australia	95 %	92 %	94 %	91 %
Austria	97 %	92 %	94 %	92 %
Belgium	93 %	92 %	91 %	89 %
Brazil	94 %	89 %	91 %	81 %
Bulgaria	92 %	90 %	89 %	84 %
Canada	93 %	94 %	92 %	89 %
Chile	94 %	94 %	92 %	88 %
Colombia	95 %	93 %	93 %	91 %
Croatia	96 %	96 %	95 %	93 %
Czech Republic	96 %	95 %	95 %	91 %
Denmark	92 %	94 %	94 %	89 %
Estonia	98 %	96 %	96 %	95 %
Finland	97 %	98 %	97 %	95 %
France	90 %	89 %	91 %	89 %
Germany	90 %	84 %	**76 %**	78 %
Greece	97 %	97 %	96 %	91 %
Hungary	93 %	94 %	91 %	90 %
Iceland	97 %	97 %	93 %	93 %
Indonesia	93 %	92 %	89 %	93 %
Ireland	95 %	95 %	96 %	93 %
Israel	81 %	87 %	88 %	88 %
Italy	97 %	97 %	96 %	93 %
Japan	90 %	90 %	89 %	88 %
Jordan	79 %	86 %	78 %	79 %
Korea (South)	98 %	97 %	97 %	96 %
Latvia	95 %	93 %	93 %	91 %
Lithuania	96 %	93 %	94 %	87 %
Luxembourg	90 %	90 %	90 %	87 %
Mexico	95 %	95 %	95 %	94 %
Montenegro	87 %	90 %	84 %	82 %
Netherlands	95 %	94 %	93 %	94 %
New Zealand	90 %	88 %	87 %	86 %
Norway	93 %	95 %	93 %	92 %
Poland	96 %	95 %	94 %	94 %

Portugal	96 %	97 %	94 %	94 %
Qatar	**58 %**	**80 %**	81 %	83 %
Romania	93 %	94 %	90 %	83 %
Russia	97 %	97 %	96 %	89 %
Slovak Republic	95 %	95 %	92 %	87 %
Slovenia	97 %	95 %	96 %	95 %
Spain	96 %	96 %	97 %	93 %
Sweden	95 %	93 %	92 %	89 %
Switzerland	97 %	95 %	96 %	92 %
Taiwan	94 %	94 %	95 %	88 %
Thailand	95 %	90 %	87 %	**77 %**
Tunisia	95 %	95 %	89 %	79 %
Turkey	91 %	87 %	87 %	87 %
United Kingdom	88 %	88 %	89 %	81 %
United States of America	93 %	94 %	93 %	91 %
Uruguay	95 %	94 %	93 %	90 %
AVERAGE	93 %	93 %	92 %	89 %

Note. The smallest value in each cycle is indicated in bold.

Appendix E:

Percent of the Sample with Data on Home Possessions Scores

	2006	2009	2012	2015
Australia	99.5 %	98.9 %	98.7 %	97.8 %
Austria	99.8 %	99.0 %	99.6 %	99.6 %
Belgium	99.8 %	99.4 %	98.8 %	99.0 %
Brazil	99.4 %	94.3 %	98.4 %	94.8 %
Bulgaria	98.0 %	99.0 %	99.0 %	97.8 %
Canada	96.5 %	98.2 %	98.2 %	97.4 %
Chile	98.0 %	98.2 %	98.8 %	99.0 %
Colombia	99.7 %	99.4 %	99.2 %	98.4 %
Croatia	99.9 %	99.9 %	99.8 %	98.7 %
Czech Republic	99.8 %	99.9 %	99.6 %	99.0 %
Denmark	99.4 %	99.0 %	99.0 %	98.6 %
Estonia	99.8 %	99.8 %	99.2 %	98.7 %
Finland	99.9 %	99.6 %	99.1 %	98.9 %
France	98.6 %	99.6 %	98.5 %	98.2 %
Germany	96.8 %	**93.0 %**	**84.9 %**	**87.4 %**
Greece	99.9 %	99.7 %	99.6 %	99.3 %
Hungary	99.6 %	99.9 %	99.2 %	98.7 %
Iceland	99.0 %	98.8 %	97.3 %	97.5 %
Indonesia	100.0 %	99.8 %	99.9 %	99.8 %
Ireland	98.4 %	97.9 %	99.4 %	99.2 %
Israel	96.0 %	98.1 %	97.1 %	98.8 %
Italy	99.6 %	99.9 %	99.5 %	98.1 %
Japan	100.0 %	99.6 %	98.7 %	99.9 %
Jordan	99.8 %	99.5 %	99.1 %	99.3 %
Korea (South)	99.9 %	99.9 %	99.9 %	99.5 %
Latvia	99.8 %	99.9 %	98.4 %	99.2 %
Lithuania	99.9 %	99.6 %	99.7 %	97.8 %
Luxembourg	99.7 %	99.7 %	99.7 %	99.5 %
Mexico	99.8 %	99.6 %	99.1 %	99.2 %
Montenegro	99.2 %	99.7 %	99.0 %	98.2 %
Netherlands	99.9 %	99.5 %	98.9 %	99.2 %
New Zealand	99.2 %	99.0 %	98.8 %	97.6 %
Norway	98.0 %	99.5 %	98.2 %	97.4 %
Poland	99.8 %	99.7 %	99.7 %	99.8 %
Portugal	99.8 %	99.7 %	98.6 %	98.8 %

Qatar	**95.5 %**	98.8 %	96.8 %	99.1 %
Romania	99.9 %	99.9 %	99.8 %	100.0 %
Russia	99.9 %	99.8 %	99.4 %	95.9 %
Slovak Republic	100.0 %	99.8 %	99.1 %	98.6 %
Slovenia	99.6 %	99.5 %	99.1 %	99.2 %
Spain	99.7 %	99.4 %	98.9 %	99.0 %
Sweden	99.3 %	99.4 %	98.3 %	98.5 %
Switzerland	99.9 %	99.8 %	99.5 %	99.5 %
Taiwan	99.9 %	99.9 %	99.7 %	99.9 %
Thailand	99.9 %	99.9 %	99.9 %	97.8 %
Tunisia	99.7 %	100.0 %	98.8 %	95.8 %
Turkey	100.0 %	99.8 %	99.3 %	99.4 %
United Kingdom	98.7 %	98.9 %	98.7 %	97.8 %
United States of America	99.4 %	99.3 %	99.2 %	99.0 %
Uruguay	99.1 %	98.6 %	99.1 %	98.4 %
AVERAGE	99.2 %	99.2 %	98.7 %	98.4 %

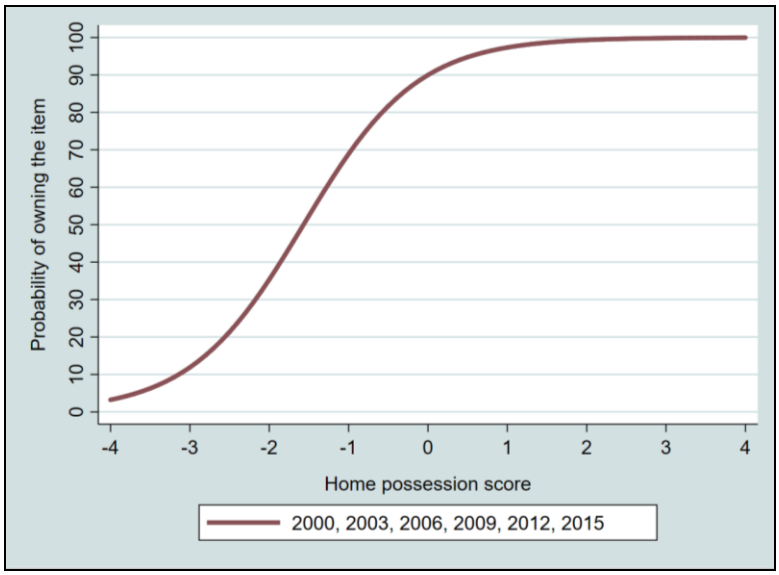
Note. The smallest value in each cycle is indicated in bold.

Appendix F:

Model-Based Item Characteristic Curve(s) for Each Item

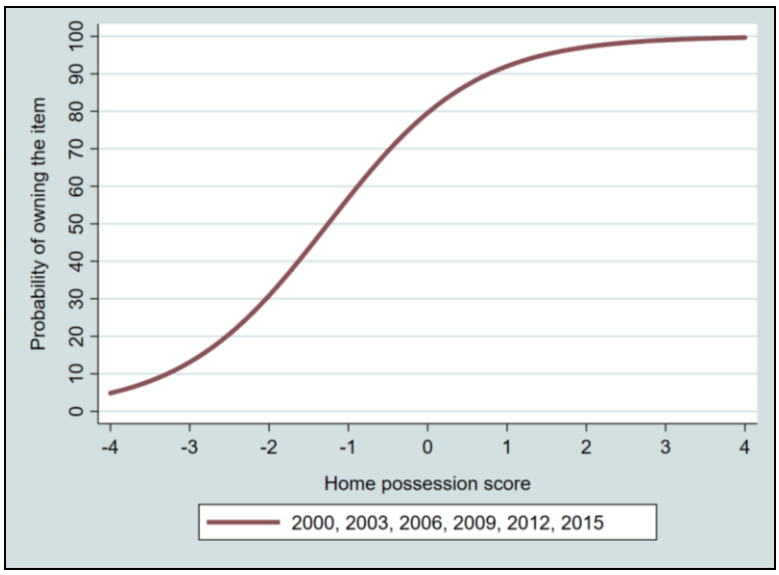
Desk

α = 0.82, β = -1.57



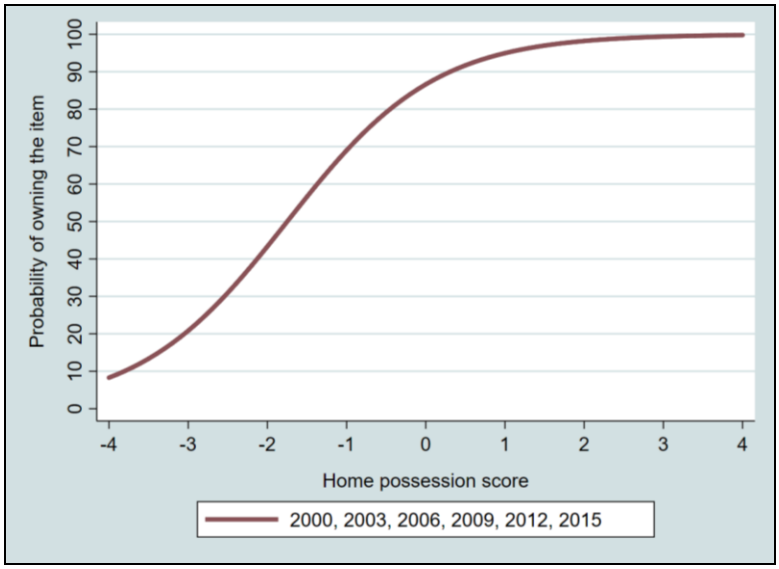
Own room

α = 0.64, β = -1.26



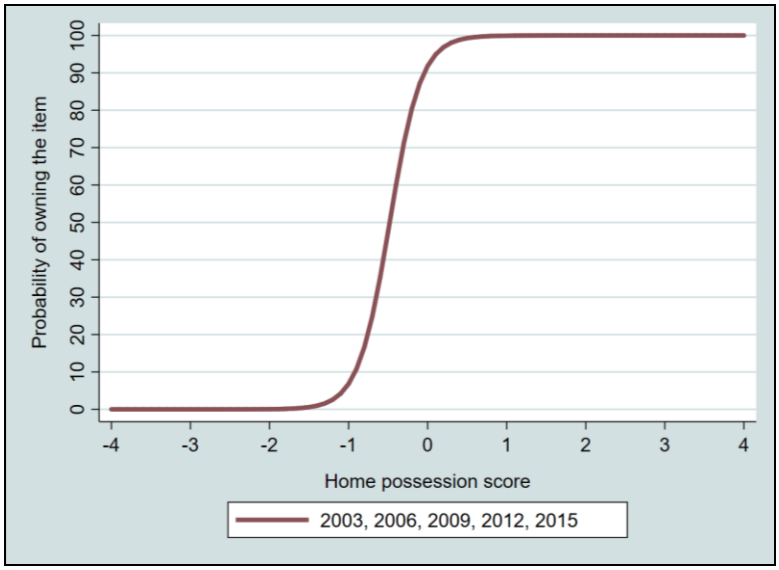
Quiet study place

α = 0.63, β = -1.75



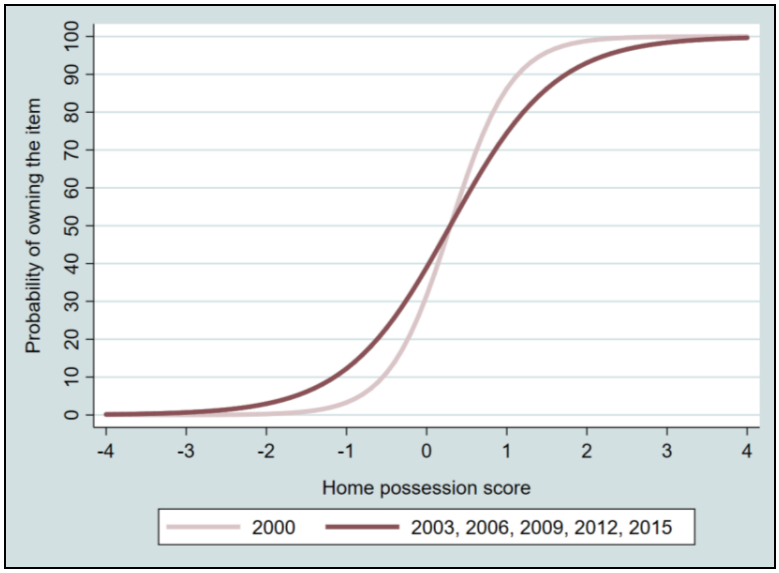
Computer (dichotomous)

α = 2.96, β = -0.48



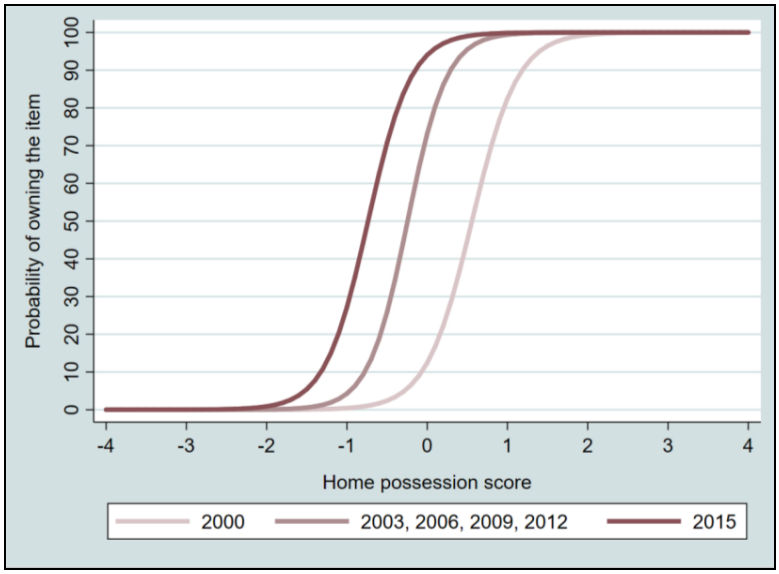
Educational software

2000: α = 1.53, β = 0.30
2003 to 2015: α = 0.89, β = 0.29



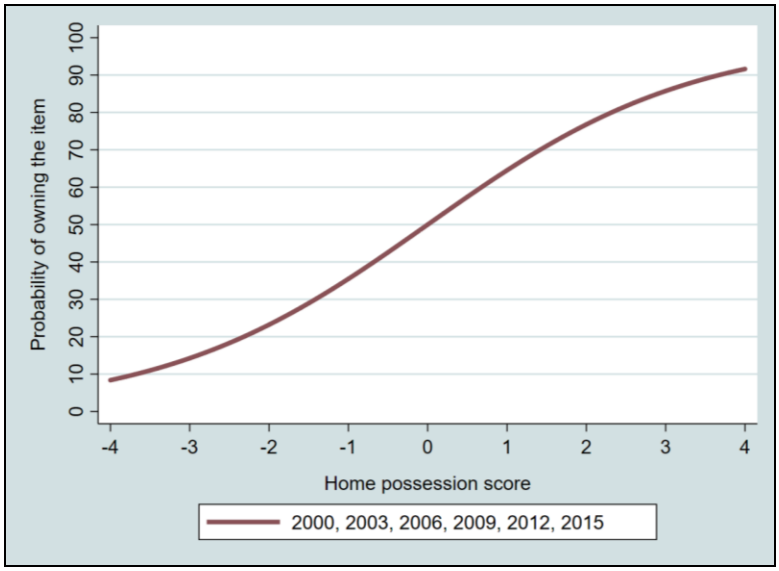
Internet

2000: α = 2.06, β = 0.56
2003 to 2012: α = 2.42, β = -0.24
2015: α = 2.20, β = -0.74



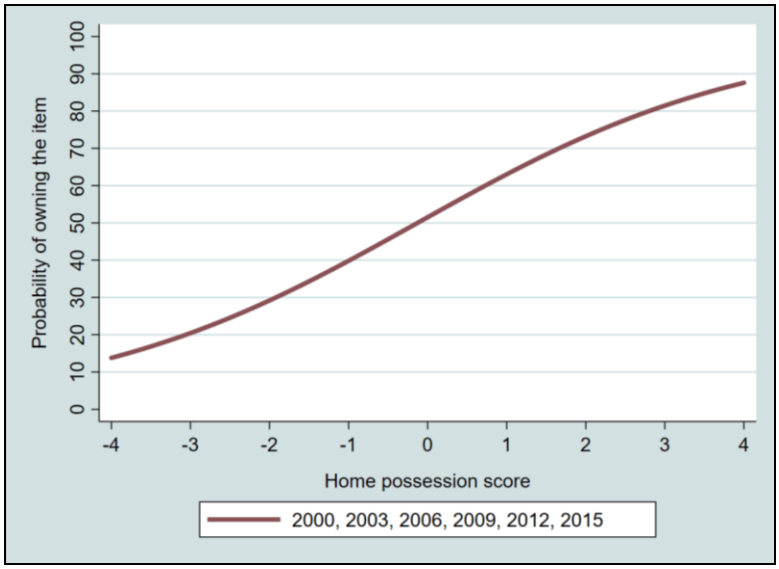
Classic literature

α = 0.35, β = 0.00



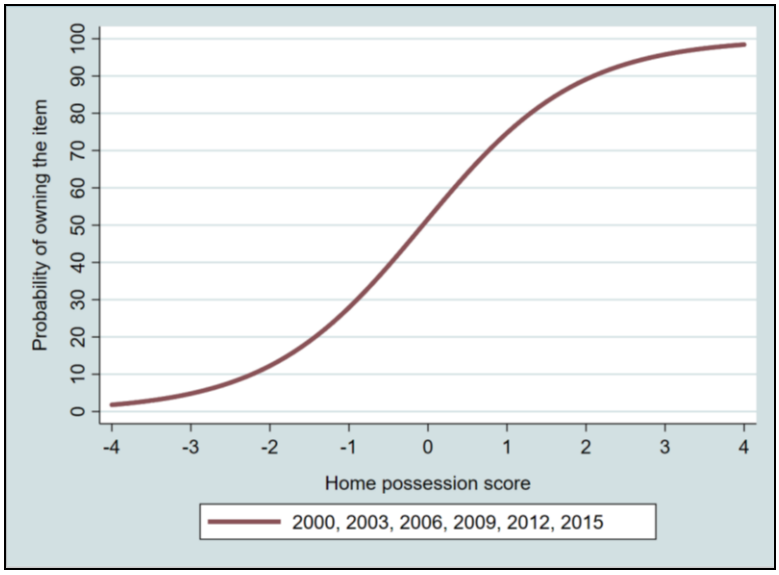
Poetry books

α = 0.28, β = -0.13



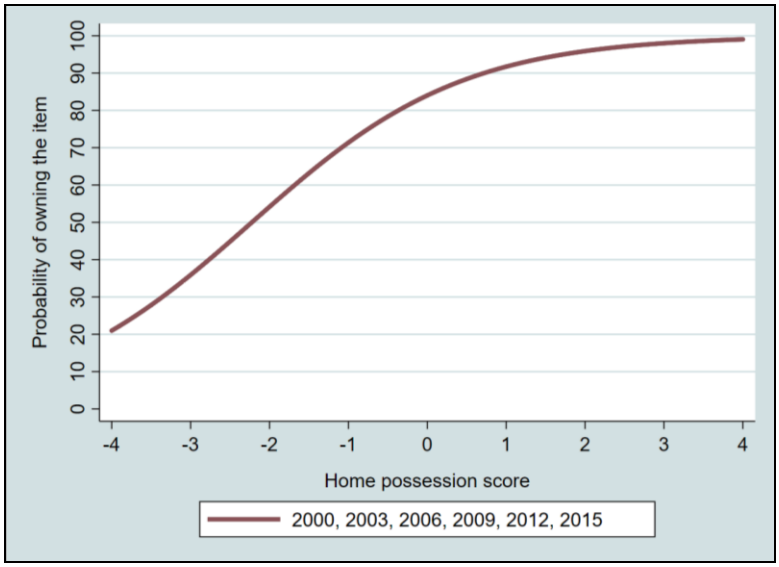
Artwork

α = 0.60, β = -0.07



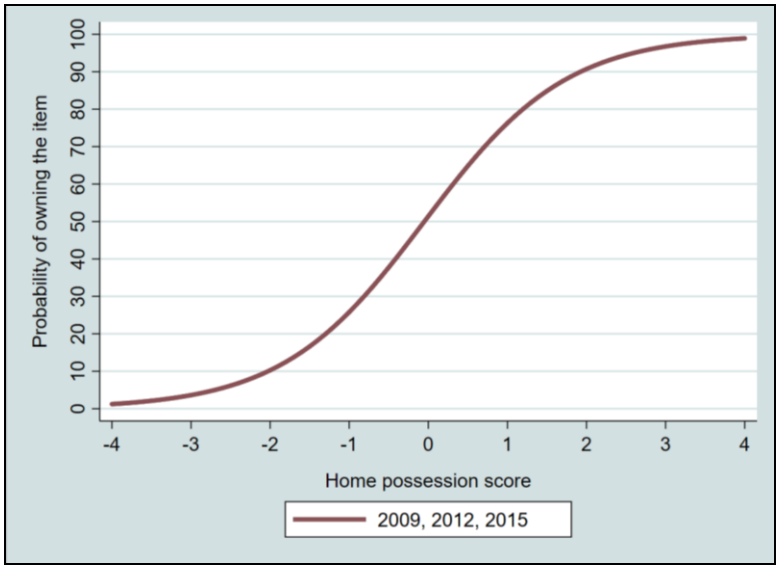
School books

α = 0.44, β = -2.22



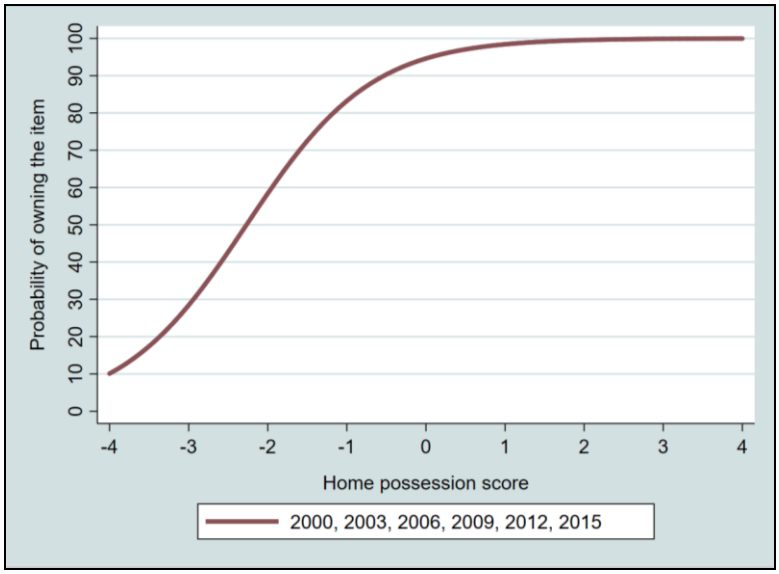
Reference books

α = 0.65, β = -0.05



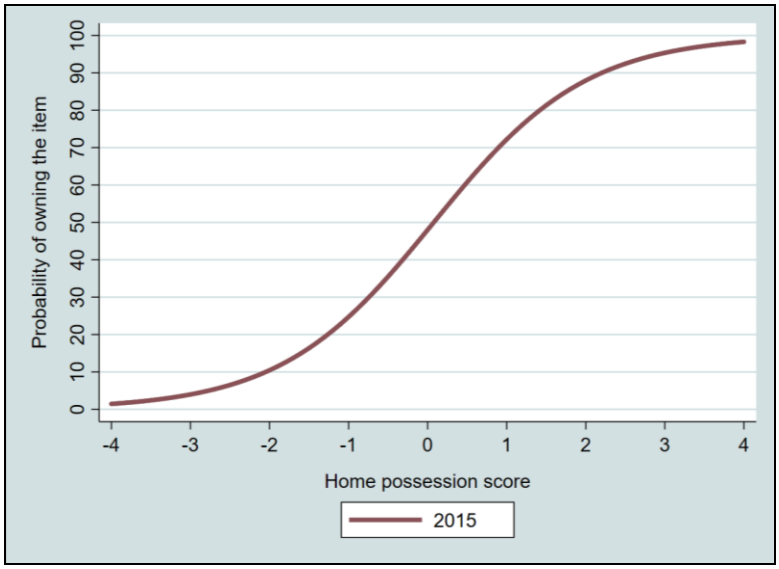
Dictionary

α = 0.74, β = -2.27



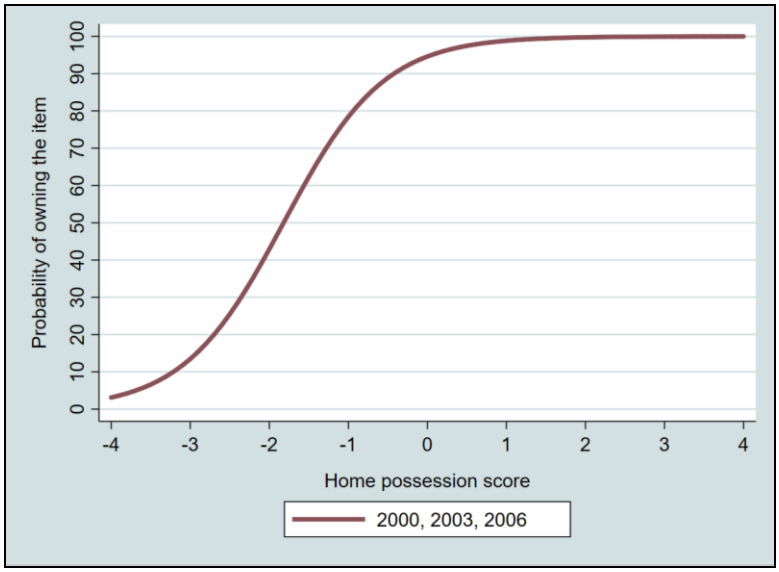
Books on culture

α = 0.61, β = 0.08



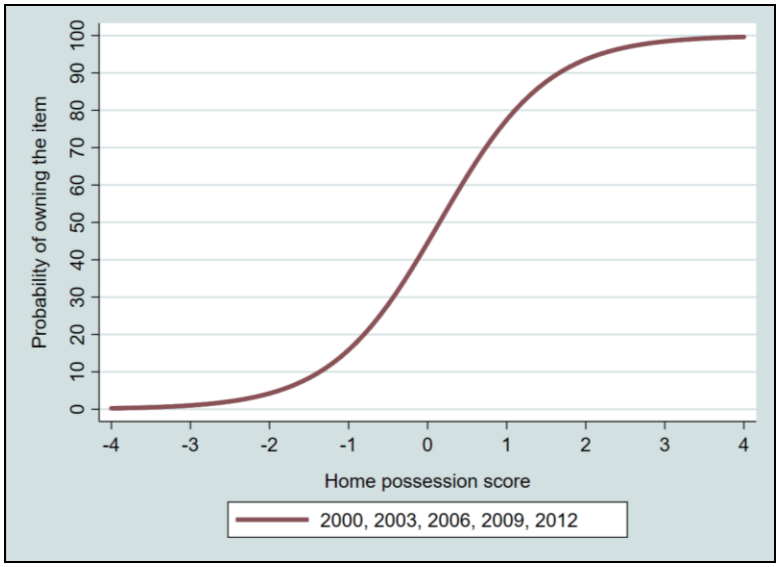
Calculator

α = 0.93, β = -1.82



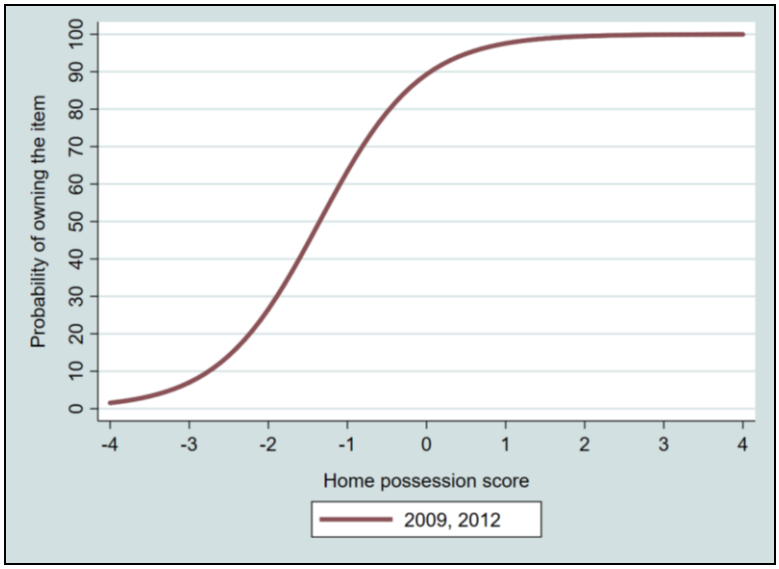
Dishwasher

α = 0.85, β = 0.15



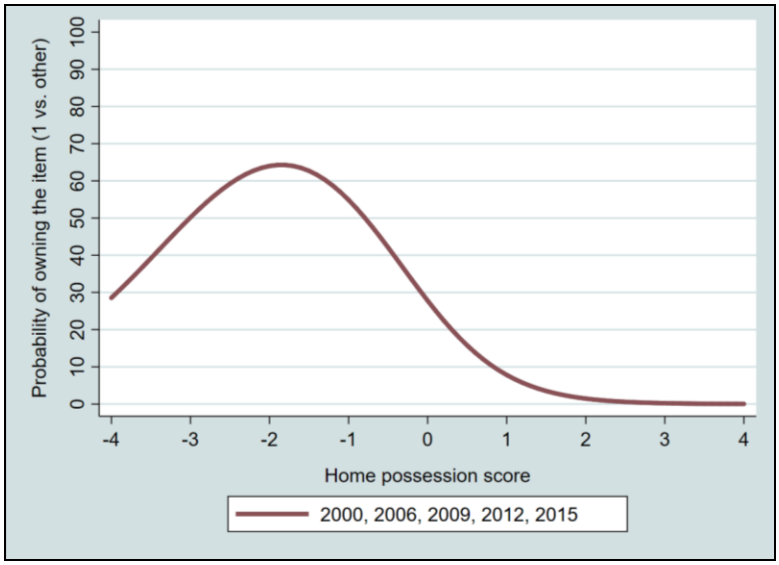
DVD player

α = 0.92, β = -1.35



TV

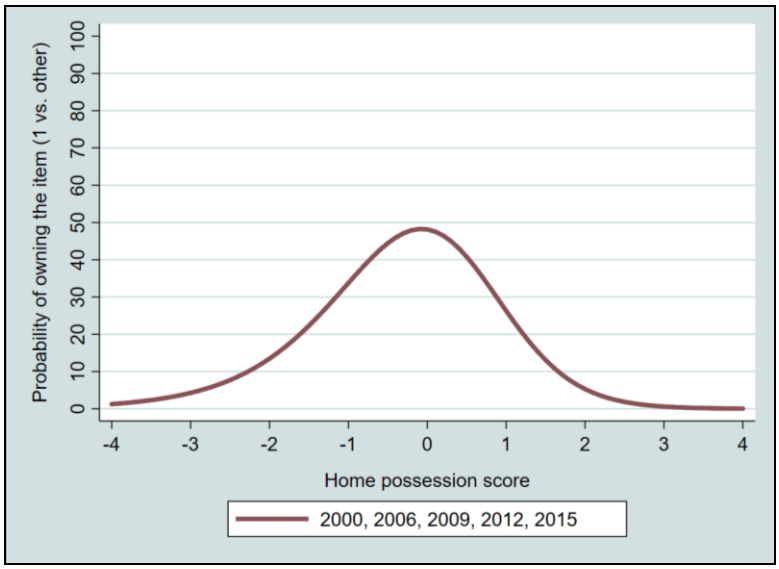
α = 0.59, β = -1.05



Note. For simplicity, only one category response curve is shown.

Car

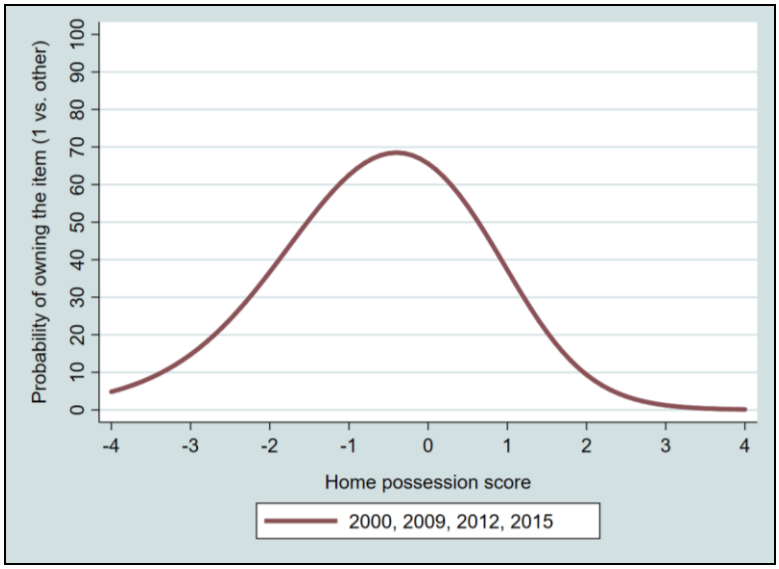
α = 0.74, β = 0.50



Note. For simplicity, only one category response curve is shown.

Bathroom

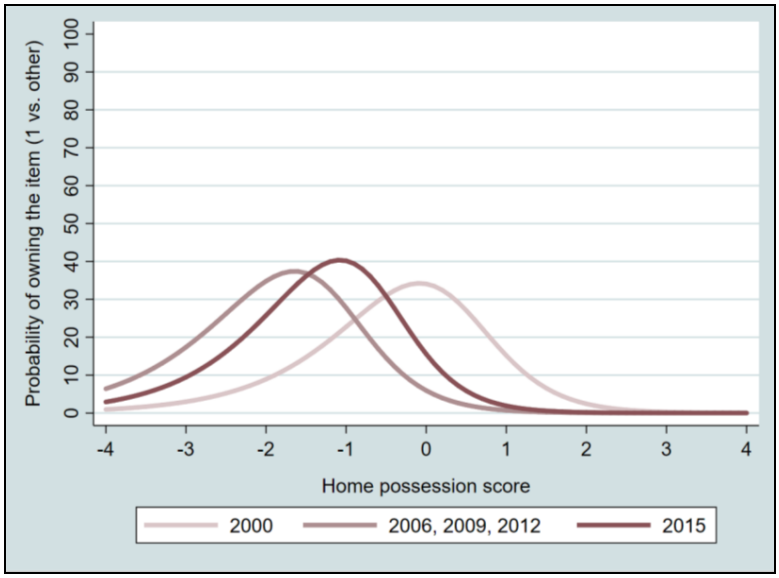
α = 0.72, β = 0.33



Note. For simplicity, only one category response curve is shown.

Cellphone

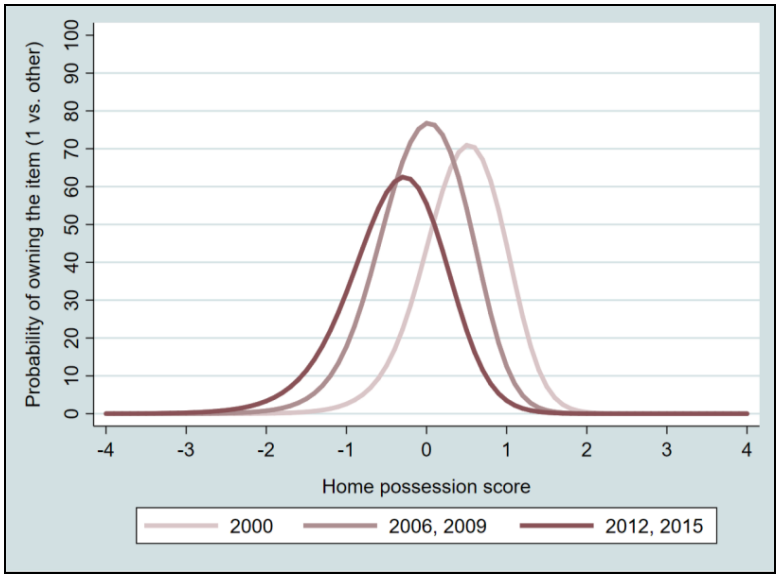
2000: α = 0.68, β = 0.32
2006 to 2012: α = 0.67, β = -1.31
2015: α = 0.73, β = -0.76



Note. For simplicity, only one category response curve is shown.

Computer (polytomous)

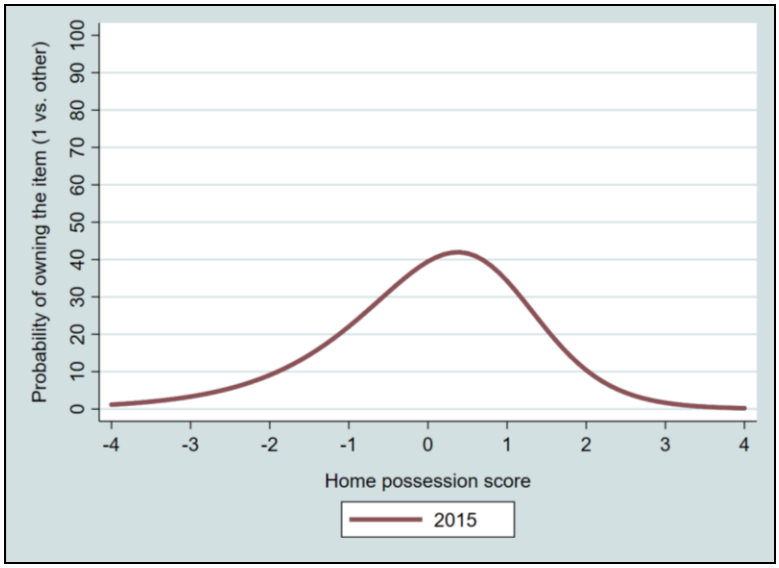
2000: α = 2.00, β = 0.82
2006 to 2009: α = 1.95, β = 0.36
2012 to 2015: α = 1.55, β = 0.05



Note. For simplicity, only one category response curve is shown.

Tablet

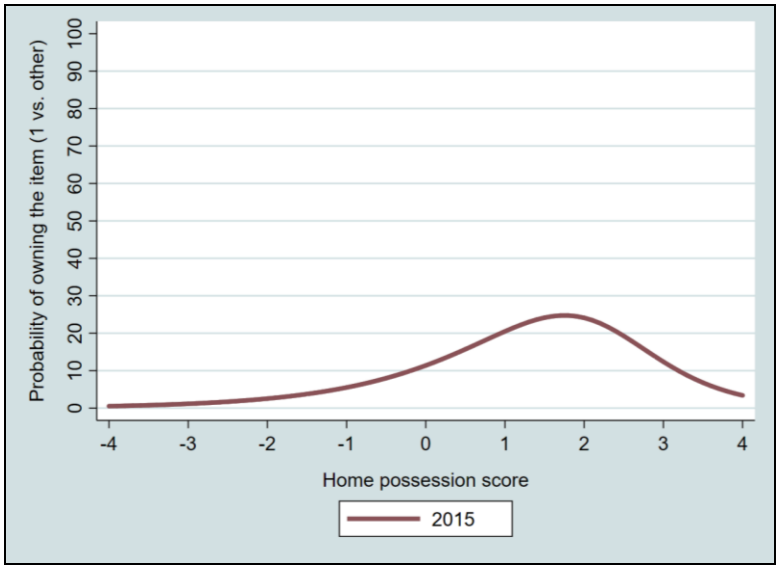
α = 0.63, β = 0.82



Note. For simplicity, only one category response curve is shown.

Ebook reader

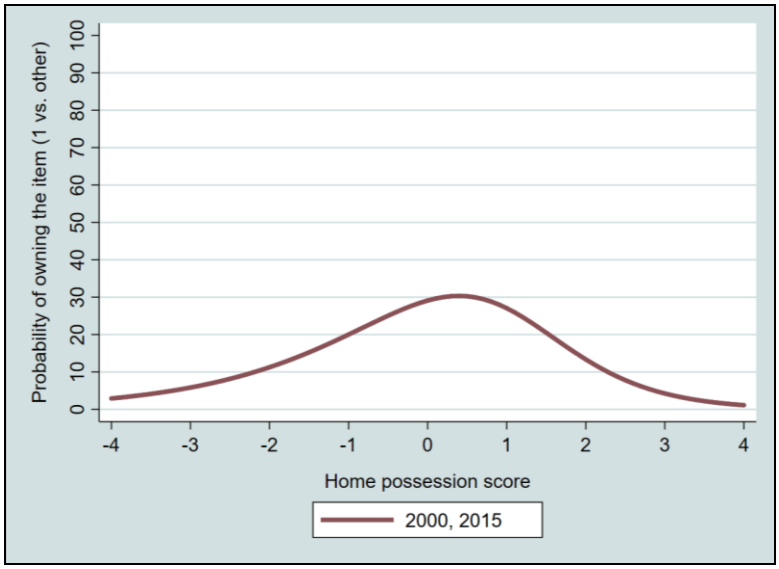
α = 0.48, β = 2.18



Note. For simplicity, only one category response curve is shown.

Instrument

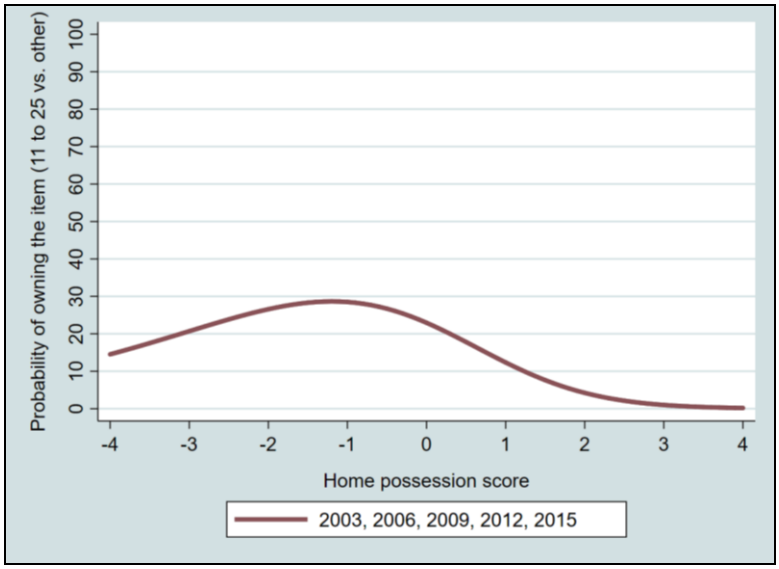
α = 0.43, β = 0.92



Note. For simplicity, only one category response curve is shown.

Books

α = 0.29, β = 0.70



Note. For simplicity, only one category response curve is shown.

Appendix G:

Percent of Country-by-Language Groups that Required Unique Item Parameters, by Item

Item (Cycle)	# of country-by-language groups that administered the item	# of country-by-language groups that required unique item parameters for the item	% of country-by-language groups that required unique item parameters for the item
Desk (2000 to 2015)	96	16	17 %
Own room (2000 to 2015)	96	9	9 %
Quiet study place (2000 to 2015)	96	6	6 %
Computer (2003 to 2015)	96	4	4 %
Ed software			
Ed software (2000)	32	5	16 %
Ed software (2003 to 2015)	96	27	28 %
Internet			
Internet (2000)	32	2	6 %
Internet (2003 to 2012)	88	3	3 %
Internet (2015)	81	2	2 %
Classic literature (2000 to 2015)	96	60	63 %
Poetry books (2000 to 2015)	96	54	56 %
Artwork (2000 to 2015)	96	17	18 %
School books (2000 to 2015)	96	13	14 %
Reference books (2009 to 2015)	96	24	25 %
Dictionary (2000 to 2015)	96	6	6 %
Books on culture (2015)	81	19	23 %
Calculator (2000 to 2006)	72	6	8 %
Dishwasher (2000 to 2012)	88	47	53 %
DVD player (2009, 2012)	88	8	9 %
TV (2000, 2006 to 2015) *	96	53	55 %
Car (2000, 2006 to 2015) *	96	38	40 %
Bathroom (2000, 2009 to 2015) *	96	66	69 %
Cellphone *			
Cellphone (2000) *	32	11	34 %
Cellphone (2006 to 2012) *	88	13	15 %
Cellphone (2015) *	81	12	15 %
Computer *			

Computer (2000) *	32	5	16 %
Computer (2006 to 2009) *	87	28	32 %
Computer (2012 to 2015) *	89	30	34 %
Tablet (2015) *	81	14	17 %
Ebook reader (2015) *	81	9	11 %
Instrument (2000, 2015) *	84	9	11 %
Books (2003 to 2015) *	96	13	14 %

Note. Polytomous items are indicated with an asterisk.

Appendix H:

Percent of Items that Required Unique Item Parameters, by Country-by-Language Group

Country (Language)	# of items adminis -tered to the group	# of items that required unique item parameters	% of items that required unique item parameters
Albania (Albanian)	27	7	26 %
Algeria (Arabic)	22	4	18 %
Argentina (Spanish)	27	4	15 %
Australia (English)	32	8	25 %
Austria (German, English)	32	3	9 %
Azerbaijan (Azerbaijani) *	21	7	33 %
Azerbaijan (Russian) *	21	10	48 %
Belgium (Dutch, German) *	28	5	18 %
Belgium (French) *	28	4	14 %
Brazil (Portuguese)	32	3	9 %
Bulgaria (Bulgarian)	32	6	19 %
Canada (English) *	28	9	32 %
Canada (French) *	28	8	29 %
Chile (Spanish)	32	4	13 %
Colombia (Spanish)	28	8	29 %
Costa Rica (Spanish)	27	5	19 %
Croatia (Croatian)	28	1	4 %
Czech Republic (Czech)	32	5	16 %
Denmark (Danish)	32	10	31 %
Dominican Republic (Spanish)	22	8	36 %
Estonia (Estonian) *	28	3	11 %
Estonia (Russian) *	28	7	25 %
Finland (Finnish) *	28	7	25 %
Finland (Swedish) *	28	6	21 %
France (French)	32	6	19 %
Georgia (Georgian, Azerbaijani, Russian)	27	10	37 %
Germany (German)	32	2	6 %
Greece (Greek)	32	2	6 %
Hungary (Hungarian)	32	3	9 %
Iceland (Icelandic)	32	1	3 %
Indonesia (Indonesian)	32	10	31 %

Ireland (English, Irish)	32	7	22 %
Israel (Hebrew, English, French, Spanish) *	28	5	18 %
Israel (Arabic) *	28	3	11 %
Italy (Italian, German, Slovenian)	32	7	22 %
Japan (Japanese)	32	13	41 %
Jordan (Arabi)	28	8	29 %
Kazakhstan (Kazakh) *	21	10	48 %
Kazakhstan (Russian) *	21	9	43 %
Korea, South (Korean)	32	12	38 %
Kosovo (Albanian)	22	9	41 %
Kyrgyzstan (Kyrgyz) *	21	11	52 %
Kyrgyzstan (Russian) *	21	11	52 %
Kyrgyzstan (Uzbek) *	21	13	62 %
Latvia (Latvian) *	28	6	21 %
Latvia (Russian) *	28	7	25 %
Lebanon (French) *	22	5	23 %
Lebanon (English) *	22	3	14 %
Liechtenstein (German)	27	5	19 %
Lithuania (Lithuanian, Russian, Polish)	28	5	18 %
Luxembourg (German, English) *	28	2	7 %
Luxembourg (French) *	28	4	14 %
Macedonia (Macedonian, Turkish) *	22	7	32 %
Macedonia (Albanian) *	22	4	18 %
Malaysia (Malay) *	21	6	29 %
Malaysia (English) *	21	6	29 %
Malta (English)	27	4	15 %
Mauritius (English)	20	4	20 %
Mexico (Spanish)	32	3	9 %
Moldova (Romanian) *	27	8	30 %
Moldova (Russian) *	27	9	33 %
Montenegro (Montenegrin, Albanian)	28	8	29 %
Netherlands (Dutch)	32	9	28 %
New Zealand (English)	32	4	13 %
Norway (Norwegian)	32	8	25 %
Panama (Spanish)	20	6	30 %
Peru (Spanish)	32	9	28 %
Poland (Polish)	32	5	16 %
Portugal (Portuguese)	32	3	9 %
Puerto Rico (Spanish)	22	9	41 %

Qatar (Arabic) *	28	15	54 %
Qatar (English) *	28	7	25 %
Romania (Romanian) *	28	13	44 %
Romania (Hungarian) *	28	4	14 %
Russia (Russian)	32	8	25 %
Serbia (Serbian, Hungarian, Slovak, Romanian)	22	6	27 %
Singapore (English)	27	10	37 %
Slovak Republic (Slovak) *	28	2	7 %
Slovak Republic (Hungarian) *	28	4	14 %
Slovenia (Slovenian, Italian)	28	3	11 %
Spain (Spanish, Galician, Valencian, Basque) *	28	2	7 %
Spain (Catalan) *	28	2	7 %
Sweden (Swedish, English)	32	7	22 %
Switzerland (German, Italian) *	28	7	25 %
Switzerland (French) *	28	5	18 %
Taiwan (Chinese)	28	7	25 %
Thailand (Thai)	32	7	22 %
Trinidad and Tobago (English)	27	8	30 %
Tunisia (Arabic)	28	3	11 %
Turkey (Turkish)	28	8	29 %
United Arab Emirates (Arabic) *	27	14	52 %
United Arab Emirates (English) *	27	9	33 %
United Kingdom (English, Welsh)	32	7	22 %
United States of America (English)	28	3	11 %
Uruguay (Spanish)	32	13	41 %
Vietnam (Vietnamese)	26	12	46 %

Note. Each cycle that required unique item parameters in Study 1 was counted as a separate item. Countries comprised of more than one country-by-language group are indicated with an asterisk.

Appendix I:

Item Parameters Estimated with Data from All Countries
and Data Excluding Countries with a Sizeable Minority Language Population
in PISA 2000

	Item discrimination parameter (α)			Item endorsement parameter (β)		
	All countries	Excluding countries with a sizeable minority language population	Differ-ence	All countries	Excluding countries with a sizeable minority language population	Differ-ence
Desk	0.96	0.94	-0.02	-1.21	-1.20	0.01
Own room	0.83	0.83	0.00	-0.83	-0.80	0.03
Quiet study place	0.64	0.61	-0.02	-1.97	-2.02	-0.04
Ed software	1.68	1.66	-0.02	0.27	0.28	0.01
Internet	2.22	2.25	0.02	0.51	0.51	0.00
Classic literature	0.26	0.32	0.06	-0.32	-0.27	0.05
Poetry books	0.15	0.19	0.04	-1.80	-1.28	0.52
Artwork	0.63	0.64	0.01	-0.15	-0.09	0.06
School books	0.39	0.38	-0.01	-3.37	-3.43	-0.06
Dictionary	1.17	1.19	0.01	-1.57	-1.57	0.00
Calculator	1.82	1.73	-0.09	-1.26	-1.37	-0.12
Dishwasher	1.39	1.32	-0.06	0.33	0.37	0.04
TV *	0.69	0.75	0.06	-0.73	-0.72	0.00
Car *	0.96	0.97	0.00	0.48	0.49	0.01
Bathroom *	0.80	0.74	-0.06	0.11	0.07	-0.04
Cellphone *	0.73	0.79	0.06	0.31	0.30	-0.01
Computer *	2.24	2.25	0.02	0.75	0.75	0.00
Instrument *	0.44	0.45	0.00	0.86	0.86	0.00

Note. Polytomous items are indicated with an asterisk.

Appendix J:

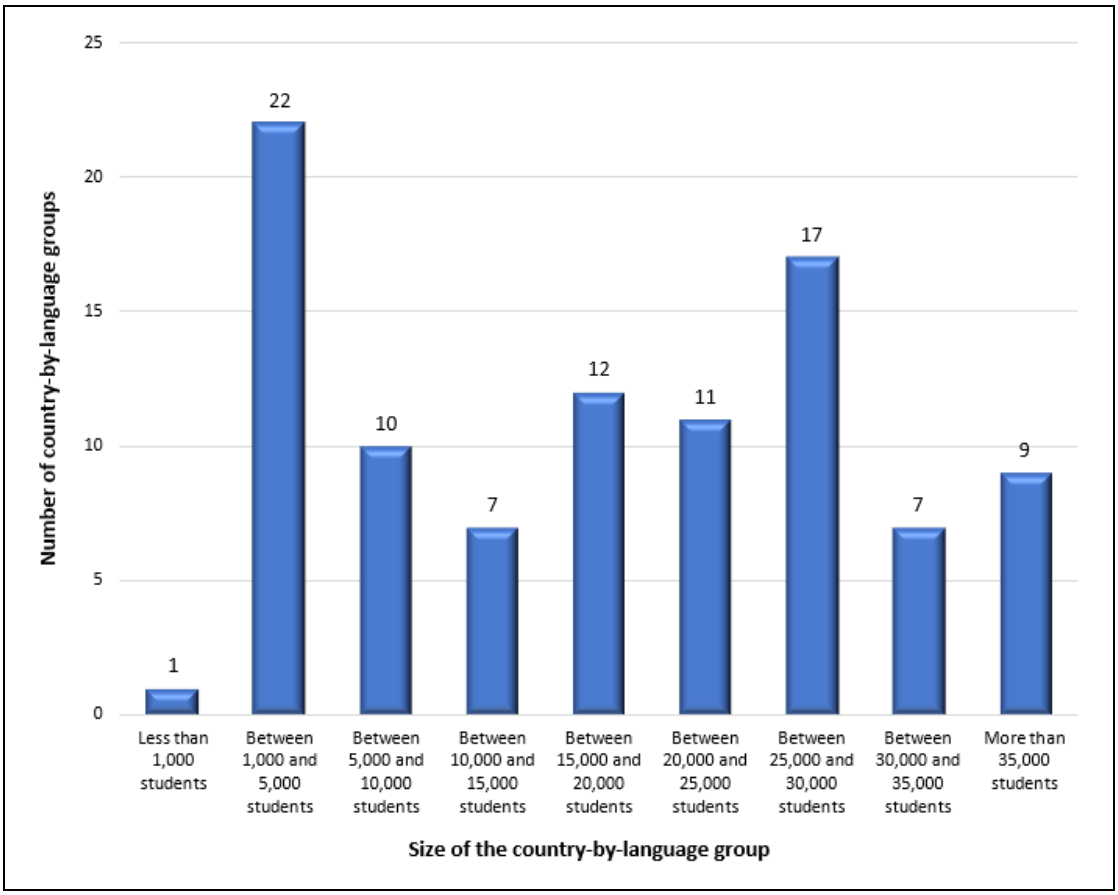
Item Parameters Estimated with Data from All Countries
and Data Excluding Countries with a Sizeable Minority Language Population
in PISA 2003

	Item discrimination parameter (α)			Item endorsement parameter (β)		
	All countries	Excluding countries with a sizeable minority language population	Differ-ence	All countries	Excluding countries with a sizeable minority language population	Differ-ence
Desk	1.07	1.04	-0.03	-0.91	-0.89	0.02
Own room	0.64	0.65	0.01	-1.00	-0.96	0.05
Quiet study place	0.71	0.67	-0.03	-0.81	-0.81	0.00
Computer	2.41	2.49	0.08	0.10	0.08	-0.02
Ed software	1.36	1.34	-0.02	0.92	0.88	-0.03
Internet	1.61	1.69	0.08	0.43	0.39	-0.04
Classic literature	0.63	0.66	0.03	0.62	0.59	-0.03
Poetry books	0.61	0.62	0.02	0.65	0.65	0.00
Artwork	0.83	0.83	0.00	0.73	0.74	0.01
School books	0.68	0.66	-0.02	-0.78	-0.85	-0.07
Dictionary	1.23	1.20	-0.04	-1.21	-1.30	-0.09
Calculator	1.05	0.99	-0.06	-1.07	-1.06	0.00
Dishwasher	0.80	0.78	-0.02	0.50	0.49	0.00
Books *	0.37	0.39	0.01	0.73	0.74	0.00

Note. Polytomous items are indicated with an asterisk.

Appendix K

Histogram of the Sample Size of the Country-by-Language Groups



REFERENCES

About PISA. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/

Akyol, Ş. P., Krishna, K., & Wang, J. (2018). *Taking PISA seriously: How accurate are low stakes exams?* (Working Paper No. 24930). Cambridge, MA: National Bureau of Economic Research.

Ben-Nun, P. (2008). Respondent fatigue. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 742-743). Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781412963947.n480

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Brese, F., & Mirazchiyski, P. (2013). *Measuring students' family background in large-scale international education studies* (IERI monograph series - Special issue 2). Princeton, NJ: IEA-ETS Research Institute (IERI).

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. doi: 10.1037/0033-2909.105.3.456

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research, 47*(4), 665-686. doi: 10.1177/0049124117701478

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*. 55-75. doi: 10.1146/annurev-soc-071913-043137

Davidov, E., Muthén, B., & Schmidt, P. (2018). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods & Research, 47*(4), 631-636. doi:10.1177/0049124118789708

DHS Overview. (n.d.). Retrieved from the DHS website: https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm

Duncan, O. D., Featherman, D. L. & Duncan, B. (1972). *Socioeconomic background and achievement*. New York, NY: Seminar Press.

Embretson, S. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Falkingham, J., & Namazie, C. (2002). *Measuring health and poverty: A review of approaches to identifying the poor*. London, UK: DFID Health Systems Resource Centre.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data – or tears: An application to educational enrollments in states of India. *Demography, 38*(1), 115-132. doi:10.1353/dem.2001.0003

Filmer, D., & Scott, K. (2008). *Assessing asset indices* (Policy Research Working Paper No. 4605). Washington, DC: World Bank.

Glas, C. A., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97-115). Boca Raton, FL: CRC Press.

Glas, C. A., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-95). New York, NY: Springer.

GDP per capita, PPP (current international dollars). (n.d.). Retrieved from the World Bank website: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

Human Development Index. (n.d.). Retrieved from the UNDP website: http://hdr.undp.org/en/content/human-development-index-hdi

Human development data. (n.d.) Retrieved from the UNDP website: http://hdr.undp.org/en/data#

IDB Analyzer [Computer software]. Downloaded from the IEA website: https://www.iea.nl/data

International Association for the Evaluation of Educational Achievement (IEA). (2014). *TIMSS 2015: Student questionnaire (Grade 8)*. Retrieved from the TIMSS & PIRLS website: https://timssandpirls.bc.edu/timss2015/questionnaires/downloads/T15_StuQ_IntSc_8.pdf

International Association for the Evaluation of Educational Achievement (IEA). (2015). *PIRLS 2016: Student questionnaire*. Retrieved from the TIMSS & PIRLS website: https://timssandpirls.bc.edu/pirls2016/questionnaires/downloads/P16_StuQ.pdf

ISCED 1997. (1997). Retrieved from the UNESCO website: http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524-545. doi:10.1037/met0000113

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi: 10.1007/BF02294825

Montgomery, M. R., Gragnolati, M., Burke, K. A., & Paredes, E. (2000). Measuring living standards with proxy variables. *Demography, 37*(2), 155-174. doi: 10.2307/2648118

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research, 47*(4), 687-728. doi: 10.1177/0049124117729691

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. doi: 10.1002/j.2333-8504.1992.tb01436.x

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313-335. doi:10.1037/a0026802

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 166-172. doi:10.3389/fpsyg.2014.00978

National Institute of Statistics and Informatics of Peru. (2018). *Perú: Perfil sociodemográfico*. Retrieved from the INEI website: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1539/libro.pdf

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315-333.

Organization for Economic Co-operation and Development (OECD). (2002). *PISA 2000 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2011). *How do some students overcome their socio-economic background?* (PISA in focus #5). Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2012). *PISA 2009 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2014a). *Parent questionnaire for PISA 2015*. Retrieved from the OECD website: http://www.oecd.org/pisa/data/CY6_QST_MS_PaQ_Final.pdf

Organization for Economic Co-operation and Development (OECD). (2014b). *PISA 2012 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2017). *PISA 2015 technical report*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2018). *Equity in education: Breaking down barriers to social mobility*. Paris, France: OECD. doi: 10.1787/9789264073234-en

PIRLS Overview. (n.d.). Retrieved from the NCES website: https://nces.ed.gov/surveys/pirls/

PISA 2000 list of participating countries/economies. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/pisa2000listofparticipatingcountrieseconomies.htm

PISA 2015 list of participating countries/economies. (n.d.). Retrieved from the OECD website: http://www.oecd.org/pisa/aboutpisa/pisa2000listofparticipatingcountrieseconomies.htm

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259-278. doi: 10.2304/rcie.2013.8.3.259

Rutstein, S. O. (2008). *The DHS wealth index: Approaches for rural and urban areas* (DHS Working Papers No. 60). Calverton, MD: Macro International.

Rutstein, S. O., & Johnson, K. (2004). *The DHS wealth index* (DHS Comparative Reports No. 6). Calverton, MD: ORC Macro.

Sandoval-Hernandez, A., Rutkowski, D., Matta, T., & Miranda, D. (2019). Back to the drawing board: Can we compare socioeconomic background scales? *Revista de Educación, 383,* 37-61. doi:10.4438/1988-592X-RE-2019-383-400

Shin, H. J., Khorramdel, L., Xu, X., & von Davier, M. (2017). *Multidimensional discrete latent trait models (mdltm).*

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453. doi:10.3102/00346543075003417

TIMSS overview. (n.d.). Retrieved from the NCES website: https://nces.ed.gov/timss/

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin, 133*(5), 859-883. doi: 10.1037/0033-2909.133.5.859

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development* (General Assembly resolution). Retrieved from the UN website: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2016). *Education 2030: Incheon Declaration and Framework for Action for the implementation of Sustainable Development Goal 4*. Retrieved from the UIS website: http://uis.unesco.org/sites/default/files/documents/education-2030-incheon-framework-for-action-implementation-of-sdg4-2016-en_2.pdf

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6*, 1-4. doi:10.3389/fpsyg.2015.01064

von Davier, M. (2005). Multidimensional discrete latent trait models (mdltm) [Computer software].

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*(3), 115-124. doi: 10.1027/1614-2241.3.3.115

Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning, 21*(6), 459-468. doi: 10.1093/heapol/czl029

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series, 2008*(1), 1-18. doi: 10.1002/j.2333-8504.2008.tb02113.x

Yamamoto, K. (1998). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey* (pp. 161-178), Washington, DC: National Center for Education Statistics.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Chapter 17: Scaling PIAAC cognitive data. In *Technical report of the survey of adult skills (PIAAC)* (pp. 406-438). Paris, France: OECD.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155-173.

Yang, Y., & Gustafsson, J. E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation, 10*(3), 259-288. doi: 10.1076/edre.10.3.259.30268